Universitat Ramon Llull

TESI DOCTORAL

Títol Definició d'una metodologia experimental per a l'estudi de resultats en sistemes d'aprenentatge artificial

Realitzada per Josep M. Martorell i Rodon

en el Centre Escola Tècnica i Superior d'Enginyeria Electrònica i Informàtica La Salle

i en el Departament Informàtica

Dirigida per Dr. Josep M. Garrell i Guiu

C. Claravall, 1-3
08022 Barcelona
Tel. 936 022 200
Fax 936 022 249
E-mail: urlsc@sec.url.es
www.url.es

RESUM

El treball presentat s'enmarca dins del camp d'actuació propi del Grup de Recerca en Sistemes Intel·ligents: l'aprenentatge artificial. Les grans àrees són la computació evolutiva i el raonament basat en casos, tot dirigint la recerca a problemes de classificació, diagnosi i predicció. En tots aquests camps són objecte d'estudi grans conjunts de dades, pels quals es treballen diferents tècniques que en permeten l'extracció de coneixement i l'aplicació als problemes citats. Els grans avenços en aquestes àrees (sovint en forma de nous algorismes) conviuen amb treballs molt parcials sobre les metodologies adequades per a l'avaluació d'aquestes noves propostes.

En front d'aquesta situació, la tesi que aquí es presenta proposa un nou marc general per a l'avaluació del comportament d'un conjunt d'M algorismes que, per tal de ser analitzats, són assajats sobre N problemes de prova. La tesi sosté que l'anàlisi habitual que es fa d'aquests resultats és clarament insuficient, i que degut a això les conclusions que s'exposen en els treballs publicats són sovint parcials, i en alguns casos fins i tot errònies.

El treball s'inicia amb un estudi introductori sobre les mesures que permeten expressar la bondat d'un algorisme, a través de l'assaig sobre una colecció de problemes de prova. En aquest punt, es demostra la necessitat d'un estudi previ de les propietats inherents d'aquests problemes (a partir, per exemple, de les mètriques de complexitat) si es vol assegurar la fiabilitat de les conclusions que s'obtindran.

A continuació, es defineix el marc d'aplicació de tot un conjunt de tècniques d'inferència estadística per les quals, essent aquestes prou ben conegudes, s'analitzen els factors a tenir en compte en la determinació del seu domini d'ús. La tesi proposa un protocol general per a l'estudi, des d'un punt de vista estadístic, del comportament d'un conjunt d'algorismes, incloent uns nous models gràfics que en faciliten l'anàlisi, i l'estudi detallat de les propietats inherents als problemes de prova utilitzats.

Aquest protocol determina el domini d'ús de les metodologies per a la comparació dels resultats obtinguts en cada problema. La tesi demostra, a més, com aquest domini està directament relacionat amb la capacitat d'aquesta metodologia per a determinar diferències significatives, i també amb la

seva replicabilitat.

Finalment, es proposen un conjunt de casos sobre resultats ja publicats amb anterioritat, fruit de nous algorismes desenvolupats pel nostre Grup de Recerca, molt en especial en l'aplicació del raonament basat en casos. En tots ells es mostra la correcta aplicació de les metodologies desenvolupades en els capítols anteriors, i es destaquen els errors comesos habitualment, que duen a conclusions no fiables.

RESUMEN

El trabajo presentado se enmarca dentro del campo de actuación propio del Grupo de Investigación en Sistemas Inteligentes: el aprendizaje artificial. Las grandes áreas son la computación evolutiva y el razonamiento basado en casos, dirigiendo la investigación a problemas de clasificación, diagnóstico y predicción. En todos estos campos son objeto de estudio grandes conjuntos de datos, para los cuales se trabajan diferentes técnicas que permiten la extracción de conocimiento y la aplicación a los citados problemas. Los grandes avances en estas áreas (muchas veces en forma de nuevos algoritmos) conviven con trabajos muy parciales sobre las metodologías adecuadas para la evaluación de estas nuevas propuestas.

Frente a esta situación, la tesis que aquí se presenta propone un nuevo marco general para la evaluación del comportamiento de un conjunto de M algoritmos que, para poder ser analizados, son ensayados sobre N problemas de prueba. La tesis sostiene que el análisis habitual que se hace de estos resultados es claramente insuficiente, i que debido a esto las conclusiones que se exponen en los trabajos publicados son muchas veces parciales, y en algunos casos hasta erróneas.

El trabajo se inicia con un estudio introductoria sobre las medidas que permiten expresar la bondad de un algoritmo, a través del ensayo sobre una colección de problemas de prueba. En este punto, se demuestra la necesidad de un estudio previo de las propiedades inherentes de estos problemas (a partir, por ejemplo, de las métricas de complejidad) si se quiere asegurar la fiabilidad de las conclusiones que se obtendrán.

A continuación, se define el marco de aplicación de todo un conjunto de técnicas de inferencia estadística para las cuales, siendo éstas bien conocidas, se analizan los factores a tener en cuenta en la determinación de su dominio de uso. La tesis propone un protocolo general para el estudio, desde un punto de vista estadístico, del comportamiento de un conjunto de algoritmos, incluyendo unos nuevos modelos gráficos que facilitan su análisis, y el estudio detallado de las propiedades inherentes a los problemas de prueba utilizados.

Este protocolo determina el dominio de uso de las metodologías para la comparación de resultados obtenidos en cada problema. La tesis demuestra,

además, como este dominio está directamente relacionado con la capacidad de esta metodología para determinar diferencias significativas, y también con su replicabilidad.

Finalmente, se proponen un conjunto de casos sobre resultados ya publicados con anterioridad, fruto de nuevos algoritmos desarrollados por nuestro Grupo de Investigación, muy en especial en la aplicación del razonamiento basado en casos. En todos ellos se muestra la correcta aplicación de las metodologías desarrolladas en los capítulos anteriores, y se destacan los errores cometidos habitualmente, que llevan a conclusiones no fiables.

Índex

Índex de figures

Índex de taules

Part I

Introducció i plantejament del treball

Capítol 1

Introducció

"Determining a suitable classifier for a given problem
is however still more an art than science"
http : //en.wikipedia.org/wiki/Artificial_intelligence

1.1 Context

En l'àrea de coneixement relacionada amb el *machine learning* un article o comunicació científica acostuma a seguir el següent esquema: en primer lloc, un nou algorisme d'aprenentatge (o alguna variant d'un altre ja existent) és presentat, tot justificant-ne la seva necessitat i els elements conceptuals que l'han permès desenvolupar. És obvi (de vegades fins i tot ni s'explicita) que aquesta nova proposta desitja aportar una millora respecte les alternatives existents fins aquell moment, ja sigui en la bondat amb què desenvolupa la tasca que té encomanada (que determinarà la qualitat dels resultats), ja sigui en el temps que necessita per dur-la a terme, etc.

A continuació, aquesta nova proposta és avaluada a partir d'un conjunt de problemes de prova, construits *ad hoc* per a l'assaig o bé provinents dels respositoris on es troben els problemes d'ús habitual per aquestes comeses (el més habitual és l'*UCI repository*, [3]). Les antigues opcions algorísmiques també són avaluades sobre aquests mateixos problemes per tal de tenir uns resultats amb què comparar la nova proposta.

Finalment, tots aquests resultats obtinguts són comparats, amb l'objectiu

de justificar perquè el nou algorisme proposat és una bona opció: millora en la precisió de la tasca objectiu, reducció de l'error, reducció en el temps de càlcul, etc. Aquest plantejament acostuma a tenir diversos punts febles perquè, a l'hora d'escollir les tècniques aplicades a cada fase (estimació, comparació, conclusions), difícilment s'estudia si el context compleix les condicions exigides per la metodologia experimental escollida i, per tant, fins a quin punt les conclusions que s'extrauran seran fiables.

Potser on això és més crític és en les comparacions, que es duen a terme a partir de test[1] d'inferència estadística. Massa sovint, aquest ús dels test no va acompanyat de l'estudi previ que justifiqui el perquè de la seva aplicació. En d'altres casos, simplement s'aplica fora del seu domini d'ús, i les conclusions trobades no tenen realment validesa, o bé són només parcials.

Cal dir que, en els darrers anys, s'observa una creixent inquietud per la validació estadística dels resultats publicats i de les metodologies utilitzades per a la comparació, així com per la correcta determinació dels rangs de validesa de les conclusions obtingudes. Alguns autors, com Demsar ([9]), ho atribueixen a la maduresa de l'àrea, a l'augment de les aplicacions sobre problemes reals (i, per tant, l'evident necessitat de justificar estrictament les propostes presentades) i la facilitat per a l'experimentació que implica disposar de força eines i repositoris per a assajar nous algorismes (aquí es podria destacar el respositori UCI abans mencionat o eines com el programari WEKA ([10]) o l'entorn conegut com a KEEL (Knowledge Extraction Evolutionary Learning, [11])).

Vista aquesta realitat, la voluntat d'aquest treball és estudiar, a nivell teòric i en la seva aplicació sobre problemes amb què s'ha trobat el nostre Grup de Recerca[2], una metodologia global que permeti una anàlisi acurada

[1] En tot el treball s'utilitzarà aquest terme per referir-se, indistintament, al singular i al plural. A banda de la influènci de l'anglès, cal reconèixer que la construcció del plural en la nostra llengua no hi ajuda: *testos* porta irremeiablement a pensar en objectes ceràmics, i *tests* es fa realment de difícil pronúncia, especialment si es repeteix cada poques línies. Ambdós formes són acceptades a la nostra llengua, però es mantindrà el *test* en tots els casos: l'autor demana disculpes, d'avançat, per aquesta llicència.

[2] El Grup de Recerca en Sistemes Intel·ligents està adscrit a l'Escola Superior d'Enginyeria Informàtica la Salle de la Universitat Ramon Llull. El grup centra la seva activitat al voltant de l' Aprenentatge Artificial, aplicat a problemes de classificació, predicció i diagnosi. Les principals tècniques comprenen la Computació Evolutiva (Algorismes Genètics, Programació Genètica, Sistemes Classificadors) i el Raonament Basat en Casos, tot i que també es tracten altres tècniques de l'aprenentatge artificial, com les Xarxes Neuronals i l' Aprenentatge Inductiu. L'activitat del grup i la seva consolidació en la temàtica ha estat reconeguda per la Generalitat de Catalunya (2002 SGR-00155, 2005 SGR-00302).

del comportament d'un nou algorisme d'aprenentatge respecte un conjunt d'algorismes prèviament coneguts. Una proposta global de metodologia rigorosa seria molt útil per a la major part de comunitat investigadora de l'àrea, a l'hora que permetria que es concentressin millor en els algorismes pròpiament.

1.2 Motivació

És realment necessari un treball amb aquest enfocament? La inquietud comentada en els paràgrafs anteriors ha anat acompanyada de publicacions amb una certa voluntat d'estudi global de la qüestió, com les de Martin i Hirschberg sobre l'avaluació de la bondat ([12]), la de Dietterich sobre la potència de test en comparacions entre dos algorismes ([13]) o la del mateix Demsar ([9]), que fa un repàs sobre diferents metodologies de comparació amb dos i més algorismes.

Tot i això, aquest darrer autor mostra que el seguiment de les propostes que s'han realitzat per part del personal investigador de l'àmbit és molt escàs: l'estudi que ell mateix realitza sobre les tècniques utilitzades per a la comparació de resultats entre algorismes és prou indicatiu. Segons aquest estudi, la pràctica totalitat dels articles publicats a l'*International Conference on Machine Learning* (ICML) entre els anys 1999 i 2003 s'allunyen de les metodologies adequades per a l'anàlisi de la bondat de la proposta que realitzen en el seu article, i fins i tot són molt poques aquelles que intenten fer un mínim anàlisi estadístic complert de les dades obtingudes. Fent una ullada a les contribucions al propi ICML el darrer 2006 ([14]), o a d'altres congressos internacionals de referència ([15], [16]), s'obté un panorama similar.

Quines són les causes d'aquest poc seguiment, que en les etapes inicials del treball s'observà també en les publicacions del propi Grup de Recerca, que ha animat a la seva realització? Fàcilment es poden identificar un conjunt de factors que ho provoquen: d'entre els que caldria destacar l'absència d'un treball conegut que estudiï com dur a terme, matemàticament parlant, totes les etapes d'un treball en l'àmbit del *machine learning*, que venen a ser les identificades a l'inici de l'apartat anterior. Tot plegat ha provocat que la

Actualment, alguns dels projectes més destacats en què participa són el *Keel II: Modelos Evolutivos de extracción de reglas. Aplicación a Data Mining. Complejidad de los Problemas de Clasificación y Diseño de Experimentos* (TIN2005-08386-C05-04), i el *MID-CBR: Un marco integrador para el desarrollo de sistemas de razonamiento basado en casos* (TIN2006-15140-C03-03).

majoria de treballs optin per un *10-fold cross-validation* ([17]) per estimar la bondat de l'algorisme, i apliquin després alguns t-test ([18]) sobre diverses comparacions per a discutir els resultats, sense més anàlisi ni de les condicions per aplicar-ho, ni del control del nivell de confiança, ni del plantejament de les hipòtesis, etc.

Els estudis publicats fins al moment no cobreixen totes les parts del procés, i tampoc ho fan amb una total completitud. A més, les tècniques que s'hi exposen no duen a una sistemàtica precisa i clara, aplicable en tots els casos en funció de les característiques del problema. Tampoc no es prenen en consideració qüestions que posteriorment es desvetllaran com a transcendents (com, per exemple, les propietats inherents als problemes de prova utilitzats per a l'estimació de la bondat sels algorismes), i es centren només en aquelles etapes i tècniques que ja són més conegudes: el *cross-validation*, el t-test, etc.

Finalment, la majoria dels estudis realitzats es situen en un dels extrems possibles: o bé són extremadament teòrics i es troben molt allunyats de la terminologia i les aplicacions habituals de l'àmbit, la qual cosa n'acostuma a descartar la seva adopció amb facilitat per part de la comunitat del *machine learning*, o bé es basen estrictament en resultats experimentals i, per tant, la seva extrapolació a una sistemàtica general és difícilment vàlida.

Un dels exemples més clars de l'allunyament entre aquests "dos móns" es trobarà en els capítols 5 i 8 d'aquest treball, a l'hora d'estudiar la viabilitat de l'aplicació d'un determinat test d'inferència estadística per a comprovar una hipòtesi. Tal i com s'estudiarà als apartats 6.2.3 i 7.3.2, la possibilitat d'aplicar un determinat test a un problema depèn de que es compleixin tot un conjunt de condicions sobre els resultats obtinguts. Aquestes condicions són perfectament conegudes des d'un punt de vista teòric (veure, per exemple, el text de Sheskin [19]) i la seva comprovació admet una certa relaxació en alguns problemes, com es plantejarà en els estudis del seu domini d'ús, en aquests mateixos apartats.

Com a alternativa a l'estudi d'aquestes condicions, en molts casos s'estudien conceptes com la "potència" o la "replicabilitat" d'un test (veure capítol 8), per tal d'avaluar la idoneïtat de la seva aplicació en aquell problema concret ([20]). Aquestes són unes magnituds que en els darrers anys han aparegut en alguns articles publicats per la comunitat del *machine learning*, per una raó ben simple: la seva definició, com es veurà, és bastant més comprensible que la de les condicions per al domini d'ús dels test, i el seu càlcul també és força més fàcil de realitzar.

No obstant això, en el citat capítol 8 es mostrarà com existeix una relació

directa entre aquests dos conceptes i el compliment de les condicions del domini d'ús d'un test, convenientment relaxades en funció del que permeti el problema: són dos enfocaments des d'àmbits diferents que porten a les mateixes conclusions, com no podria ser d'una altra manera. Representen, per tant, un clar exemple d'aquests dos extrems que de vegades són difícils de fer compatibles.

Aquesta és una de les principals motivacions del present treball: acostar aquests "dos móns", i facilitar a la comunitat del *machine learning* l'ús de metodologies avançades, aplicant-les amb total correcció i aprofitant-ne tota la seva capacitat.

Així doncs, i vistes les motivacions i alguns dels antecedents, l'objectiu d'aquest treball és doble: d'una banda, es vol procedir a un estudi que, des de la primera fins l'última etapa d'un problema de *machine learning*, indiqui de quina manera procedir per al corresponent anàlisi i quines són les alternatives possibles, el perquè de la seva utilització, etc.; d'altra banda, es vol desenvolupar aquest estudi tenint en compte els conceptes teòrics necessaris per a la justificació de les propostes, però també considerant els problemes, exemples i terminologia habitual en l'entorn del *machine learning*. Possiblement, només d'aquesta manera les conclusions que se n'extreguin seran realment aplicables a la resolució de les habituals problemàtiques que apareixen en aquest camp, i seran adoptades amb facilitat per la pròpia comunitat.

1.3 Estructuració del treball

El conjunt del treball segueix un esquema que intenta reproduir la seqüència de qüestions que apareixen en un estudi d'un nou algorisme d'aprenentatge, amb especial atenció al plantejament de les hipòtesis ("l'algorisme A és comporta millor que l'algorisme B", etc.) i la seva discussió. Aquestes etapes apareixen representades a la figura 1.1, on amb un color diferent es fa també referència als capítols d'aquest treball en què s'estudien cadascuna d'elles.

Un cop establert els antecedents, la terminologia, els algorismes i problemes de prova que s'utilitzaran per a l'estudi experimental de les propostes presentades (**capítol 2**), la segona part del treball s'inicia amb el **capítol 3**, en què es planteja la resposta a la primera de les preguntes que cal respondre per tal d'estudiar un nou algorisme d'aprenentatge: de quina manera es pot avaluar la seva bondat, i quina o quines magnituds ho poden fer?

La tasca que tingui encomanada l'algorisme determinarà, en bona mesura, la magnitud que serà d'utilitat per avaluar la seva bondat. L'àmbit de coneix-

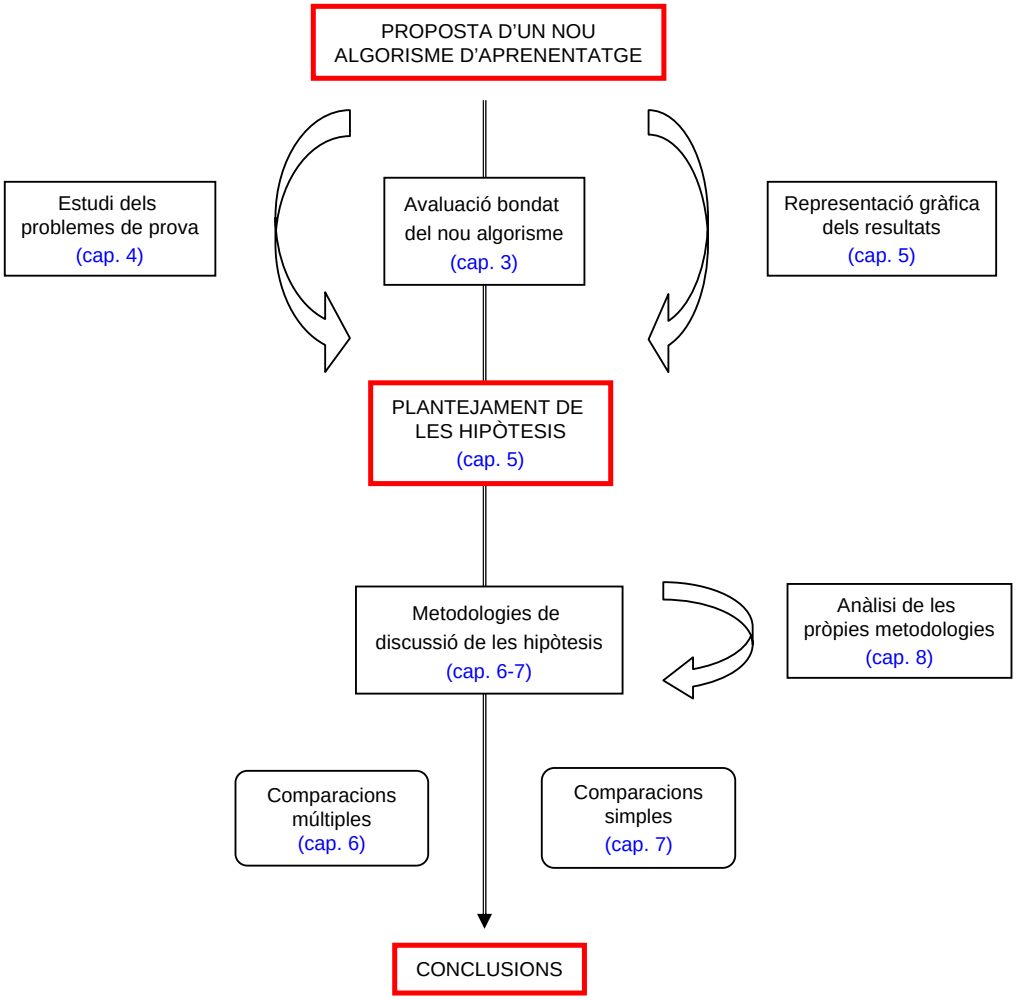

Figura 1.1: Esquema que reprodueix les etapes de l'estudi d'un nou algorisme d'aprenentatge, incloent-hi les operacions per obtenir una magnitud de la seva bondat i les comparacions amb les alternatives existents per valorar-ne el guany. En blau es mostren els capítols d'aquest treball en què es discuteix cada etapa.

ement i les aplicacions en què treballa el nostre Grup de Recerca, porten a què habitualment es treballi amb sistemes classificadors. En aquest cas, l'error o precisió en la classificació acostumarà a ser l'escollida per a discutir sobre la bondat. L'estudi sobre les fonts d'error de l'algorisme que es fa a la primera part d'aquest capítol 3 té validesa per qualsevol magnitud utilitzada com indicador de bondat. A més, a partir dels resultats publicats a [5], s'estudia la possibilitat de rebutjar una classificació sistemàtica, tot reduint els casos efectivament classificats però augmentant molt més la precisió de l'algorisme, cosa que pot ser extremadament útil en casos en què l'existència d'un error tingui un cost molt elevat.

A continuació s'analitza la qüestió de com avaluar la bondat d'un algorisme. Conceptualment, la pròpia pregunta hauria de ser sorprenent: no deixa de ser-ho que, en un àmbit *madur* i en el qual s'ha avançat tant com aquest, calgui encara dedicar un cert temps a discutir quin és el sistema de mesura de l'experiment. Sense tenir clares les condicions experimentals de mesura, sembla difícil imaginar com es pot procedir a l'anàlisi de les hipòtesis plantejades, per exemple.

En el cas que ens ocupa, es pot començar la discussió suposant que un algorisme A és generat amb l'objectiu de millorar l'algorisme B en una tasca determinada (per exemple, la classificació) sobre un espai X format pels problemes existents (allò que caldrà classificar). La bondat de cada un d'ells sobre aquest espai X no es pot calcular, doncs possiblement X serà infinit o, com a mínim, massa gran per a intentar-ho. D'aquesta manera, el problema es redueix a l'avaluació d'ambdós algorismes sobre un sub-espai $S \subset X$, format pels problemes de prova de què es disposin.

A partir d'aquestes consideracions, cal canviar el verb "avaluar" pel d'"estimar", en el sentit estadístic del terme, doncs el que es farà serà elaborar algun procés de A i B sobre S que estimi quina és aquesta bondat. D'aquí es dedueix que el valor obtingut tindrà un cert biaix respecte el real, que buscarem que sigui mínim. A més, com els mètodes proposats tenen tots ells un cert component aleatori, el resultat serà diferent en cada prova, i per tant es desitjarà també una variància mínima per a l'estimador. Aquestes dues qüestions vindran, a més, condicionades pel fet que l'aprenentatge de l'algorisme i l'estimació de la bondat es farà sobre el mateix problema de prova.

A aquests temes es dedica la segona part del capítol 3: l'estudi d'aquesta qüestió es conclou amb un conjunt de consideracions sobre cada un dels estimadors habitualment utilitzats (a partir dels amplis treballs que ja s'han realitzat sobre aquesta qüestió, veure l'apartat 3.4) que, tot i acceptant la no

existència d'un estimador amb comportament òptim en tots els problemes possibles, acaba proposant el *m-times 10-fold cross-validation* amb estratificació com l'opció òptima en la majoria dels casos.

Un cop estimada la bondat dels algorismes del problema (habitualment, de la nova proposta i de les alternatives existents), cal plantejar les hipòtesis d'acord amb el que desenvolupa en el primer apartat del **capítol 5**, en què s'estudia com estructurar en cada cas la qüestió a discutir, i quin ha de ser el control a efectuar sobre el nivell de significança de la hipòtesi: en tant que es tracta d'un problema d'inferència estadística, no té sentit una conclusió que rebutgi o accepti una hipòtesi sense més, sinó una conclusió que la rebutgi o accepti amb un determinat percentatge de confiança en aquesta conclusió.

Per tal de plantejar correctament aquestes hipòtesis, però, cal estudiar abans una altra qüestió: l'efecte que tenen les propietats inherents dels problemes de prova utilitzats sobre la pròpia conclusió obtinguda. En aquest punt el treball introdueix resultats obtinguts de l'aplicació de les dues grans famílies de tècniques amb què el grup treballa: la computació evolutiva i el reaonament basat en casos (CBR). La relació entre els resultats i les propietats dels problemes de prova, en concret, s'analitzen a partir d'unes variants del CBR definida en clusteritzar la memòria de casos (introduïdes a [21], i que donen lloc al conegut com a SOMCBR).

De fet, ha estat amb l'estudi en profunditat de les citades variants que s'han vinculat els resultats obtinguts amb propietats inherents dels problemes de prova, com s'ha publicat a [2]. L'estudi dels valors de certes mètriques de complexitat aplicades a aquests problemes ([22], [23]) ha permès determinar que per a un plantejament correcte de les hipòtesis a discutir cal estudiar abans aquestes magnituds, doncs d'altra manera és possible que un gran nombre de problemes de prova "amagui" possibles diferències significatives entre els algorismes, en una tipologia determinada de problemes. Totes aquestes qüestions es plantegen al **capítol 4**, incloent-hi la definició d'una nova mètrica de complexitat que porta a l'estructuració d'un mapa on els problemes es poden agrupar per diferents regions de complexitat, i analitzar de manera més simple quins són els efectes d'aquests sobre les conclusions obtingudes.

Un altre aspecte a tenir en compte és la representació gràfica dels resultats, tema tractat al segon apartat del capítol 4. A partir de les dades obtingudes de bondat, amb l'estimador escollit per al problema, i potser també d'altres mesures de qualitat (la qual cosa ens portaria a un problema multivariant) es poden construir esquemes gràfics que permetin una primera visualització dels resultats de les noves propostes i de les ja existents. Aquesta primera imatge, molt més fàcil d'analitzar que les habituals i interminables

taules de resultats (especialment per problemes multivariants), dóna també idea de com plantejar les hipòtesis a discutir (què comparar respecte què, per exemple), i permet visualitzar l'efecte de les propietats inherents dels problemes de prova (a partir del concepte de regions de complexitat que s'haurà definit al capítol 4). D'altra banda, un cop discutides les hipòtesis amb les corresponents metodologies, aquests esquemes gràfics permetran una visualització de les conclusions a les quals s'arriba.

Un cop plantejades les hipòtesis, es desenvolupa la tercera part del treball amb els **capítols 6** i **7**, en els quals es presenten les diverses metodologies per a la discussió d'aquestes hipòtesis. L'element bàsic que fa una primera distinció de les opcions disponibles és el fet que es tracti d'una comparació simple (comparació entre 2 algorismes, capítol 6) o múltiple (més de dos algorismes per comparar, capítol 7). Per ambdós casos, el capítol inclou una proposta de protocol d'actuació en funció del compliment de totes les condicions que determinen el domini d'ús d'un test paramètric d'inferència estadística.

Dins de cada cas, la primera decisió a prendre és la utilització d'un test paramètric o d'un test no-paramètric: de l'anàlisi de les condicions teòriques per a la fiabilitat del resultat, i la seva particularització en els problemes d'aprenentatge, en surten unes tècniques per a determinar el seu domini d'ús que determinaran els protocols que es proposen per a la discussió de les hipòtesis plantejades (veure els esquemes dels apartats 6.2.3 i 7.5).

En aquest punt cal aclarir el perquè contínuament es parla d'hipòtesis, en plural, quan aparentment només caldria discutir una sola hipòtesi (tipus "l'algorisme nou té un comportament equivalent als ja existents" o bé "l'algorisme nou té un millor comportament que els ja existents"). En el cas dels problemes de comparació múltiple, els criteris establerts al capítol 8 porten a determinar, en primer lloc, una hipòtesi global a discutir (tipus "tots els algorismes es comporten de manera equivalent"). Només en el cas que aquesta hipòtesi es pugui rebutjar, es poden plantejar tot un conjunt d'hipòtesis d'abast menor sobre les comparacions entre algorismes, per trobar quins són aquells que provoquen el rebuig de la primera hipòtesi més global.

La quarta i última part d'aquest treball s'obre amb el **capítol 8**, on s'estudien diferents magnituds que aporten informació sobre la potència i la replicabilitat dels test d'inferència que es poden utilitzar. El primer concepte parla de la capacitat dels test de posar de relleu diferències significatives entre algorismes quan aquestes existeixin, mentre que el segon dóna una idea de l'estabilitat de la conclusió del test: en tant que procés estadístic, sobre els mateixos algorismes i els mateixos problemes de prova la conclusió pot arribar

a ser diferent, i per tant es desitjarà que les conclusions obtingudes siguin molt repetibles.

En el càlcul d'aquestes variables sobre els resultats publicats a [2] es mostra com les conclusions que s'obtenen són del tot coherents amb la comprovació del domini d'ús dels test, desenvolupat als capítols 6 i 7: en aquells casos en què l'aplicació d'un test paramètric sigui possible, la seva potència i replicabilitat serà major que les d'un test no paramètric, en tant que la informació del problema que utilitza per a discutir les hipòtesis és major. En canvi, en aquells casos en què el test paramètric no estigui dins el seu domini d'ús (cosa que obligarà a l'ús d'un test no paramètric), aquest mostra una potència i replicabilitat menor.

Dit d'una altra manera, l'estudi d'aquestes qüestions no és estrictament necessari si, anteriorment, ja s'ha realitzat un correcte estudi de les condicions que determinen el domini d'ús de cada un dels test que es poden aplicar per al rebuig o l'acceptació de les hipòtesis plantejades.

Tot aquest desenvolupament permet procedir amb un esquema d'actuació com el de la figura 1.1 per a l'anàlisi comparatiu de nous algorismes d'aprenentatge, amb la seguretat

1. que les mesures de bondat utilitzades seran òptimes (en el sentit d'un estimador de biaix i variància mínims),

2. que les propietats inherents dels problemes de prova no amagaran conclusions sobre la bondat dels nous algorismes sobre certs grups de problemes,

3. que les hipòtesis del problema estaran correctament plantejades (amb el corresponent control sobre la confiança del resultat),

4. i que la metodologia utilitzada per a la discussió de les hipòtesis estarà dins el seu domini d'aplicació i, per tant, que no només els resultats que se'n deriven seran fiables (al nivell de confiança determinat abans), sinó que de totes les opcions s'utilitzarà aquella amb major potència i replicabilitat.

El treball finalitza amb l'anàlisi de tot un conjunt de casos ja publicats pel nostre Grups de Recerca (**capítol 9**), on es mostra l'aplicació de totes les tècniques desenvolupades al llarg dels capítols, i es confronta amb les conclusions publicades en el seu moment, sovint errònies o parcials, com una manera d'explicitar les mancances detectades en les metodologies habituals.

Aquest treball, doncs, proposa una reflexió que durà a plantejar una metodologia experimental per a estudiar els resultats obtinguts de sistemes d'aprenentatge artifical, i poder d'aquesta manera discutir sobre la bondat, en termes comparatius, d'un conjunt d'aquests sistemes. Aquesta tesi pretén ser una nova aportació en un entorn en què, com ja s'ha dit, és difícil trobar una anàlisi global, en quant al procés, i ben justificada des d'un punt de vista teòric.

Capítol 2

Antecedents i plantejament

"There is however no golden standard for making such comparisons
and the tests performed often have dubious statistical foundations
and lead to unwarranted and unverified conclusions"
J. Demsar, [9]

En el capítol anterior s'ha introduït quina és la voluntat d'aquest treball: proposar tot un conjunt de metodologies experimentals que permetin una acurada anàlisi del comportament d'un nou algorisme, respecte els algorismes existents que tenen la mateixa tasca encomanada. En aquest capítol es posaran les bases del treball que es desenvoluparà a partir del capítol 3, tot definint els problemes que es tractaran i la terminologia utilitzada, i relacionant-ho amb el treball previ realitzat i l'estat de l'art.

2.1 Definició i terminologia

El cas més habitual dels algorismes d'aprenentatge que s'estudiaran són els sistemes classificadors, és a dir, aquells que com a tasca principal tenen encomanada la classificació d'un conjunt de casos en diverses classes. El problema es pot definir de la següent manera ([12]): donat un conjunt finit d'elements d'una població (també anomenats instàncies), descrits pel valor d'alguns atributs i ja classificats, l'objectiu és inferir un mecanisme per predir la classe de qualsevol altre membre de la citada població, a partir només dels seus atributs.

Al llarg del treball apareixeran exemples de diferents estratègies per a desenvolupar aquesta tasca, com són el raonament basat en casos (CBR, i les diferents variants de SOMCBR introduïdes en els darrers anys pel nostre Grup de Recerca, [21]), els algorismes genètics (representació ADI del coneixement sobre sistemes LCS, per exemple, [4]) o altres sistemes classificadors com les xarxes neuronals o els desenvolupats en arbres de decisió (C4.5, IB1,...). Tots ells són algorismes que aprenen de manera inductiva (és a dir, intentant extreure regles o patrons de coneixement de massius conjunts de dades) i dins l'àmbit de l'aprenentatge supervisat: l'algorisme genera una funció que *mapeja* les entrades a sortides desitjades. És a dir, proposa una funció o un conjunt de regles que atorguen una classe (sortida) a cada un dels elements de la població (entrada).

Un problema de classificació com el descrit es diu que és *ill-posed* ([24], [25]), perquè no és un problema ben plantejat en el sentit determinat per Hadamard ([26]), pel qual un problema ho és quan:

1. Existeix una solució.
2. La solució és única.
3. La solució depèn *contínuament* de les dades, en el sentit donat per una topologia *raonable*.

Dit d'una altra manera, la quantitat d'informació de què es disposa no és mai suficient com per a determinar totalment una solució única i, per tant, el propi classificador es basa en tot un conjunt de suposicions que porta directament al conegut com a principi del *no-free-lunch* ([27]): cap estratègia per inferir un nou classificador serà millor que qualsevol de les altres per a tots els problemes existents. A més, sempre caldrà realitzar algunes suposicions, que implicaran un cert error en la solució final.

Totes les estratègies porten a la construcció d'un classificador per al qual es plantegen immediatament tot un conjunt de preguntes: quina és la seva bondat, com estimar-la de manera precisa, com de millor és aquesta respecte la d'altres classificadors, i amb quina confiança es pot respondre aquesta darrera qüestió. Aquest treball presenta un conjunt de metodologies per a respondre a aquestes preguntes, tot utilitzant la terminologia descrita en els paràgrafs següents.

En primer lloc, es considera una població com un conjunt d'elements que es voldran classificar, que estan descrits per tot un conjunt d'atributs i que pertanyen a alguna de les classes possibles. Aquests atributs poden prendre

valors numèrics, ordinals o nominals, i a partir d'ells s'inferirà un algorisme classificador que construeixi una relació entre els valors dels atributs i les classes possibles.

La bondat de l'algorisme vindrà determinada per la capacitat que tindrà de fer la tasca encomanada, a partir d'aquesta relació generada. En el cas d'un sistema classificador, la seva bondat vindrà avaluada per magnituds com l'error o la precisió en la classificació (a banda d'altres com el seu cost computacional, per exemple). L'estimació d'aquest valor es farà a partir de l'assaig del classificador sobre uns problemes de prova: conjunts d'elements de la població, per als quals es coneixen el valor dels atributs i la classe a la qual pertanyen.

Un cop coneguda la bondat del classificador, es voldrà realitzar la comparació amb d'altres ja coneguts. Aquesta comparació es farà plantejant en primer lloc una hipòtesi: per exemple, la que es coneixerà com a hipòtesi nul·la i que consisteix en considerar que el nou algorisme té la mateixa bondat (fa igual de bé la seva feina) que els ja coneguts. En contraposició a aquesta, hi haurà el que es coneixerà com la hipòtesi alternativa i que s'acceptarà en cas de rebuig de la hipòtesi nul·la.[1]

La discussió de les hipòtesis es duu a terme amb els test d'inferència estadística. Entre els algorismes comparats hi haurà una certa diferència en la magnitud que avalua la bondat, i caldrà discutir si aquesta diferència és fruit només de l'atzar (acceptació de la hipòtesi nul·la), o bé és que *realment* hi ha una diferència de comportament entre els algorismes (rebuig de la hipòtesi nul·la), i per això els resultats obtinguts sobre la col·lecció de problemes de prova són diferents. La decisió es prendrà a partir dels test, que retornen un valor per a l'estadístic que determina la probabilitat que sigui certa la primera hipòtesi: per tant, caldrà determinar el nivell de confiança que s'està disposat a utilitzar, la qual cosa determinarà l'acceptació o no de la hipòtesi nul·la.

Els test que s'utilitzaran per a la discussió d'aquestes hipòtesis seran objecte d'un profund estudi, i vindran determinats per si la comparació és simple (la hipòtesi nul·la implica dues mesures de bondat) o bé múltiple (la hipòtesi nul·la n'implica més de dues). El primer cas inclou la comparació per parelles (un algorisme respecte un altre) o la comparació complexa (un algorisme respecte una combinació d'altres). En tots els casos, es podrà optar per test paramètrics, amb moltes restriccions per al seu ús però elevada potència[2], i els no paramètrics, que pràcticament no necessiten del compli-

[1] Tota aquesta terminologia serà desenvolupada més a fons al capítol 5.

[2] Aquest concepte serà definit amb exactitud al capítol 8, en què s'introduirà en relació a la conveniència d'aplicar un o altre test d'inferència. A banda de les qüestions concretes

ment de cap condició sobre les dades, però que són menys capaços de trobar diferències existents.

2.2 Estat de l'art

En el capítol introductori s'ha exposat quina era la motivació del treball, justificada pel fet que els treballs previs existents no cobrien, en totalitat, les qüestions que es plantegen en aquest tipus de problemes. Aquestes qüestions es poden agrupar en tres àmbits: l'avaluació de la bondat dels algorismes, les metodologies de comparació del comportament dels algorismes (i l'avaluació de les pròpies metodologies), i l'estudi dels problemes de prova.

En el primer, un treball a destacar és el que van publicar Martin i Hirschberg el 1996 ([12]), on estudiaven el biaix i la variància dels estimadors de bondat dels sistemes classificadors, a través de l'aplicació de diverses variants d'aquests (classificadors lineals, classificadors basats en veïns propers, arbres de decisió) sobre grans col·leccions de problemes. Els propis autors assumeixen que l'estudi és estrictament experimental, i que ve a donar resposta a les contradiccions aparegudes en treballs anteriors que utilitzaven un nombre molt menor de problemes i d'algorismes. Precisament, uns anys abans equips com els encapçalats per Efron ([28]), Breiman ([29]), Jain ([30]) i Weiss ([31]), havien publicat diferents treballs per comparar el biaix i la variància dels estimadors que anaven desenvolupant, però tots els experiments van ser realitzats sobre problemes de prova construïts *ad-hoc* als objectius del treball, o bé sobre problemes amb un número d'elements baix.

Kohavi ([32], [33]) també realitzà, pràcticament al mateix temps, estudis més extensos i desenvolupats tenint en compte elements nous en la representació del resultat o el rang d'aplicació dels classificadors. Posteriorment a aquests treballs, la sensació és que el focus d'atenció s'ha anat desviant cap a la qüestió de les metodologies de comparació d'aquests estimadors (el segon àmbit dels comentats abans), acceptant les conclusions d'aquests dos autors, i un principi general també d'altres qüestions: cap estimador pot ser el millor en totes les circumstàncies estudiades i problemes discutits ([27], [34]).

Totes aquestes propostes es basen sempre en l'ús de l'error o la precisió com a mesures per avaluar la bondat del classificador. Posteriorment, els treballs de Provost ([1]) i Langley ([35]) han introduït la proposta d'utilitzar magnituds relacionades amb les corbes operacionals (*Receiver Operat-*

que allí es discutiran, de moment és suficient saber que és una magnitud relacionada amb la capacitat d'un test per trobar una diferència significativa, quan aquesta existeix.

ing Characteristic, ROC) per a avaluar aquesta mesura. D'aquí n'han sorgit interessants treballs ([36]) que, tot i això, no afecten les bases del que s'estudiarà en aquesta tesi i, a més, no han estat encara massa adoptats per la comunitat del ML. En les línies de futur es proposaran algunes idees per al seu ús, seguint les conclusions d'aquest treball.

Entrant al segon àmbit dels abans referits (les tècniques de comparació de mesures de la bondat), la problemàtica de la comparació simple ha estat extensament tractada, però mantenint aquella dicotomia explicada al capítol anterior: o bé l'estudi és extremadament teòric i es troba molt allunyat de la terminologia i les aplicacions habituals de l'àmbit, o bé es basa estrictament en resultats experimentals i, per tant, la seva extrapolació a una sistemàtica general és difícilment vàlida.

En primer lloc cal destacar el treball de Dietterich ([13]), en què discuteix la comparació simple a partir de diversos test de comparació combinats amb diferents estimadors de la bondat. L'estudi, seguit per Alpaydin ([37]), extreu conclusions sobre la millor opció a partir de resultats estrictament experimentals i, a més, forçant la diferència entre els classificadors manipulant-ne el seu disseny. Altres autors posteriors, com Nadeau i Bengio ([38]), Webb ([39]) o Bouckaert ([20]) proposen petites modificacions sobre els test de cara a ajustar el valor subestimat de la variància, discutit després de manera més conceptual per Bengio i Grandvalet ([40]).

El propi Bouckaert amb E. Frank ([41], [42])) introdueixen després els conceptes de potència i replicabilitat per analitzar els criteris segons els quals un test és més o menys òptim en la seva aplicació en la discussió d'una hipòtesi nul·la. Com es veurà posteriorment, la mancança de l'anàlisi és no relacionar aquests resultats (sempre estrictament experimentals) amb les condicions que determinen l'aplicabilitat dels propis test.

En l'àmbit de la comparació múltiple, Salzberg ([43]) introdueix la correcció de Bonferroni sobre el test binomial, per permetre l'extrapolació dels test de comparació simple a problemes de comparació múltiple. Ell mateix fa menció de l'anàlisi de variàncies com una possible opció per a discutir millor la hipòtesi nul·la en aquests casos, metodologia que és assajada posteriorment, junt amb alternatives no paramètriques, per autors com Hull ([44]), Vázquez ([45]) o Pizarro ([46]).

Un treball a destacar en aquesta qüestió és el de Demsar ([9]), en el qual es fa una exposició general de les tècniques existents, tant per comparacions simples com múltiples (incloent test a posteriori), i comprova els resultats per a classificadors d'ús habitual (kNN, Naive Bayes, C4.5) assajats sobre problemes del repositori UCI. Les conclusions a les que arriba, però, defugen

l'estudi de les condicions d'aplicabilitat dels test paramètrics, i acaba amb una recomanació general d'aplicació de les alternatives no paramètriques, de cara a una major simplicitat en els càlculs a desenvolupar.

Finalment, en el darrer àmbit (l'estudi de les propietats inherents dels problemes de prova), cal destacar els treballs en el camp de les mètriques de complexitat, introduïdes per T. K. Ho i M. Basu ([47]). Posteriorment, diversos treballs s'han ocupat de l'estudi del domini de competència de certs classificadors, és a dir, d'aquells problemes pels quals l'algorisme en qüestió mostra un bon comportament, i la relació que té això amb els valors de les citades mètriques. Alguns d'ells han contat també amb la participació activa de membres del nostre Grup de Recerca, com un recent volum que estudia les mètriques de complexitat en l'àmbit del *pattern recognition* ([48]).

2.3 Algorismes i problemes de prova utilitzats

Les metodologies estudiades en aquest treball responen a les qüestions plantejades anteriorment, i totes es basen en l'avaluació de la bondat d'un determinat algorisme, amb l'estimació feta a partir del seu assaig sobre una col·lecció de problemes de prova. Amb la voluntat de no reproduir els problemes que apareixen quan les dades amb què es treballa són "artificials" (han estat generades *ad hoc*), en tots els casos s'han treballat amb resultats obtinguts d'algorismes que han estat dissenyats pel propi Grup de Recerca, o bé que són d'ús habitual. Igualment, els problemes de prova sobre els quals s'han assajat aquests algorismes són sempre pertanyents a algun repositori internacional, o bé d'ús del propi Grup, provinents d'algun cas real.

Pel que fa als algorismes, cal dir en primer lloc que l'objectiu d'aquest treball no és discutir sobre el disseny pròpiament dit dels sistemes classificadors que s'utilitzen. Aquesta no és una tesi destinada a proposar noves variacions sobre algorismes ja existents, o a demostrar que un determinat algorisme és millor que un altre en determinades condicions. Aquest treball vol establir una metodologia, precisament, per a poder dur a terme aquestes comprovacions en qualsevol cas, i per això utilitza algorismes ja dissenyats, sense cap voluntat d'entrar a fons en les seves propietats.

Dit això, cal destacar que bona part dels problemes tracten diverses variacions del raonament basat en casos (CBR), seguint la línia de treball que desenvolupa en aquest tema el nostre Grup de Recerca ([21], [49], [50]). Els algorismes utilitzats en aquest treball estudien l'efecte de la clusterització de la memòria de casos aplicada al CBR (SOMCBR) en l'error de classificació,

a partir de variacions sobre els paràmetres del problema.

Una primera opció del SOMCBR ([22]) és resoldre la fase de recuperació escollint el cas amb més similitud del clúster triat (a partir d'una funció de similitud). Una segona opció és afegir en aquesta fase la informació corresponent als veïns més propers (1, 3 o 5), i realitzar la recuperació mitjançant un procés de votació o una funció de pertinença. Aquesta opció aporta sis variacions diferents del SOMCBR ([5]), i inclou tota una estratègia que tendeix a reduir els elements classificats, provocant una menor capacitat operativa de l'algorisme, però també un percentatge d'error extremadament reduït.

Una darrera opció és estudiar la clusterització de la memòria de casos d'un procés CBR en general, parametritzant les diferents possibilitats a partir del número de clústers que s'inclouen en la fase de recuperació, i el percentatge d'elements d'aquests clústers que s'hi tenen en compte ([2]). Aquest estudi permet assajar fins a 12 variacions diferents, a banda sempre del propi CBR "clàssic" (és a dir, sense clusterització de la memòria de casos).

En algun altre cas, s'ha utilitzat els resultats obtinguts de l'aplicació d'algorismes genètics. En concret, s'ha treballat amb LCS (*learning classifier system*) amb algorismes genètics sota enfocament de Pittsburgh, amb diferents variacions sobre una representació per regles ADI ([4], [51]). En total s'han assajat 5 variacions respecte l'esquema ADI original, generades per diferents combinacions de les dues millores presentades: la proporció de cada discretitzador en la població i el número d'instàncies per atribut.

Aquestes propostes han estat també comparades amb algorismes d'ús habitual, com són els arbres generats a partir del procediment conegut com a C4.5 ([52]), i mètodes d'aprenentatge basats en els veïns més propers com l'IB1 ([53]). Ambdós són algorismes basats en principis molt diferents, i per això s'utilitzen simultàniament per a comparar-hi la bondat d'altres propostes.

Tots aquests algorismes classificadors s'assagen sobre diversos problemes de prova, que es mostren a la taula 2.1. En ella s'hi pot observar un primer conjunt de problemes de prova extrets del repositori UCI ([3]), i un segon conjunt amb problemes que no estan en aquest repositori però són d'ús habitual del nostre Grup de Recerca.

A banda de les propietats més importants (instàncies, atributs, classes, etc.), ja reflectides a la taula, s'aporten també les referències d'aquells treballs en què es presenten o s'utilitzen aquests problemes, per facilitar el seu estudi més profund si així es desitja. Un primer grup format pels problemes *Biopsia*, *Mamografia* (també referenciat de vegades com a μCa), *MIAS-Birads*, *MIAS-*

	Prob. prova	**Atr.**	**Ins.**	**Clas.**	**Distribució per classes**
BA	Balance	5	625	3	B (49), L (288), R (288)
GL	Glass	10	214	5	buildfloat (70), buildnonfloat (76), vehicle (17), containers (13), tableware (9), headlamps (29)
HE	Hepatitis	20	155	2	1 (85), 2 (70)
HS	Heart-Statlog	14	270	2	absent (150), present (120)
IO	Ionosphere	35	351	2	b (126), g (225)
IR	Iris	5	150	3	iris-setosa (50), iris-versicola (50), iris-virginica (50)
PI	Pim	9	768	2	0 (500), 1 (268)
SE	Segment	20	2310	7	brickface (330), sky (330), foliage (330), cement (330), window (330), path (330), grass (330)
SO	Sonar	61	208	2	rock (97), mine (111)
TH	Thyroid	6	215	3	1 (150), 2 (35), 3 (30)
VE	Vehicle	19	846	4	0 (212), 1 (217), 2 (218), 3(199)
WA	Waveform	41	215	3	1 (150), 2 (35), 3 (30)
WD	wdbc	31	569	2	b (357), m (212)
WI	Wine	14	178	3	1 (59), 2 (71), 3 (48)
WP	wpbc	34	198	2	b (151), m (47)
WS	Wisconsin	10	699	2	benign (458), malign (241)
BI	Biopsia	25	1027	2	0 (530), 1 (497)
BP	bpa	7	311	2	1 (180), 2 (131)
CA	Mamografia	22	216	2	benign (121), malign (95)
DD	DDSM	143	501	4	b1(61), b2(185), b3(157), b4(98)
LE	Learning	6	648	5	A, B, C, D, E
M3	MIAS-3C	153	322	3	fatty(106), dense(112), glandular(104)
MB	MIAS-Birads	153	320	4	b1(128), b2(78), b3(70), b4(44)
TA	Tao	3	1888	2	black (944), white (944)

Taula 2.1: Descripció dels problemes de prova utilitzats en diferents apartats del treball. El primer grup de problemes prové del repositori UCI ([3]), i la resta no estan en aquest repositori però són d'ús habitual del nostre Grup de Recerca. De cada un d'ells s'indica l'habitual abreviació, el nom complert, el número d'atributs de cada instància (contant-hi la classe a la qual pertanyen), el número d'instàncies total, el número de classes, i la distribució de les instàncies per cada classe.

3C i *DDSM* estan relacionats amb el diagnosi de càncer de mama. Els dos primers provenen de l'Hospital Trueta de Girona ([49], [54]), mentre que els altres tres provenen de bases de dades públiques, amb el convenient estudi i preprocessat ([55], [56]).

De la resta de problemes que no formen part del repositori UCI, cal destacar el *learning*, generat a partir del comportament d'un total de 648 estudiants de l'assignatura "Programació" del primer curs d'enginyeria a la Universitat Ramon Llull, descrits per atributs relacionats amb el seu rendiment acadèmic durant el curs, i classificats en 5 classes diferents segons el seu rendiment al final del curs. El problema fou analitzat utilitzant tècniques CBR en què els pesos venien determinats per sistemes genètics ([57]).

Part II

Tractament previ de les dades

Capítol 3

Estimació de la bondat d'un algorisme

"The traditional machinery of statistical processes
is wholly unsuited to the needs of practical research (...).
Only by systematically tackling small sample problems on their merits
does it seem possible to apply accurate tests to practical data."
R. A. Fisher, "Statistical Methods for Research Workers", 1925

3.1 Plantejament

Com ja s'ha explicat anteriorment, l'objectiu final d'aquest treball és l'establiment d'una metodologia global per a la comparació del comportament d'un conjunt d'algorismes. Per a poder realitzar aquesta comparació cal tenir unes magnituds numèriques que indiquin quin és el comportament de cada algorisme, valors que a la pràctica s'obtindran a partir dels resultat de l'assaig d'aquests algorismes sobre una col·lecció de problemes de prova.

En aquest capítol s'analitzaran totes les qüestions relatives a aquestes magnituds. D'entrada, si l'objectiu és comparar el comportament d'un conjunt d'algorismes cal establir quin o quins indicadors poden aportar informació sobre això: en l'apartat 3.2 s'explicarà que s'entén per bondat d'un algorisme, i s'estudiaran breument diverses possibilitats, de vegades fins i tot no excloents.

En el moment en què es decideix comparar el comportament d'un conjunt d'algorismes, s'està pressuposant que tots ells cometen errors en la realització

de la tasca que tenen encomanada (en el cas dels exemples en aquest treball, la classificació). Les magnituds que indiquen la bondat del comportament d'aquests algorismes tenen totes a veure amb aquests errors, i per això a l'apartat 3.3 s'analitzen les diverses fonts que els poden produir: no totes les causes d'error hauran de ser considerades per igual en la comparació del comportament d'un conjunt d'algorismes.

Finalment, el capítol estudia aquella part més delicada a l'avaluació de la bondat: la pròpia estimació, el més realista possible, de la corresponent mesura d'aquesta bondat. S'exposaran les principals metodologies proposades per nombrosos autors, s'estudiaran les mancances de cada una d'elles i, al final, es resumiran algunes regles que permetran a l'usuari obtenir una mesura de bondat amb, com a mínim, coneixement de causa s'obre l'error que es pot estar cometent.

Per tant, en tant que es tracta d'avaluar un estimador d'una variable que mesurarà error o precisió, caldrà conèixer el propi error d'aquest estimador, és a dir, *l'error de l'error*. L'objectiu serà utilitzar un estimador de la bondat de l'algorisme que tingui el biaix i la variància tan petit com sigui possible: com en d'altres aspectes, la conclusió indicarà que cap estimador pot ser el millor en totes les circumstàncies estudiades ([27], [32], [34]).

3.2 Mesures de bondat

En primer lloc, cal definir que s'entén per la bondat d'un algorisme en tant que mesura del seu comportament. Si suposem que l'algorisme que s'estudia té encomanada una certa tasca (per exemple d'establir un conjunt de regles que permetin classificar els elements del problema) la seva bondat bé determinada per aquella o aquelles magnituds que permetin avaluar el compliment d'aquesta tasca (per exemple, el percentatge d'encert en la classificació de casos que no han estat utilitzats per entrenar l'algorisme en qüestió).

Habitualment, en el nostre àmbit, la bondat ve determinada per un percentatge que indica la capacitat d'encert (o d'error, el seu complementari) en el desenvolupament de la tasca que té encomanada. Ens hi referirem com a precisió de l'algorisme. Però no és aquesta l'única variable a considerar: també pot resultar interessant que, a banda d'una bona precisió en la tasca encomanada, l'algorisme mostri una elevada eficiència (és a dir, que impliqui un baix cost computacional) o competència (que tingui un bon comportament per un ampli rang de problemes).

Aquesta multiplicitat de valors que determinen la bondat d'un algorisme

porta a considerar l'existència dels problemes que es consideraran com a multivariants: en el capítol 4, per exemple, es discutirà un cas sobre unes variacions del CBR, on les variables que indiquen la bondat del seu comportament seran la precisió en la classificació, i el número d'operacions a realitzar en una certa fase de l'algorisme. En aquest cas, son dues les magnituds que donen una idea sobre la bondat del sistema classificador. També passarà això al final de l'apartat 3.3, on les magnituds seran l'error de classificació i el percentatge d'instàncies no classificades.

Siguin quines siguin aquestes magnituds, en tots els casos son identificables diferents fonts d'error en la seva estimació. En els apartats que segueixen s'estudiaran aquestes fonts, i molt àmpliament, les diferents opcions per estimar aquest error. Tenint en compte el context en què es mou aquest treball (la comparació de la bondat d'un conjunt d'algorismes), la primera condició que cal complir és que la variable o variables que s'utilitzen per avaluar aquesta bondat siguin les mateixes per tots els algorismes d'un determinat problema: no es poden comparar M variables (cada una d'elles aportant informació sobre els M algorismes del problema) si totes elles no signifiquen el mateix, tenen la mateixa definició, el mateix rang de valors possibles, etc.

En tot el treball s'utilitzarà com a mesura de bondat la precisió en la classificació, entesa com el percentatge de casos correctament classificats d'entre el total amb que s'ha assajat el comportament de l'algorisme. Aquesta magnitud, que en l'apartat 3.4 s'estudiarà com avaluar-la, tot just estimarà el valor del que realment és la precisió d'un sistema classificador ([33]): la probabilitat de classificar correctament un cas, seleccionat aleatòriament d'un univers de casos en que la distribució d'aquests coincideix amb la del conjunt de casos amb que s'ha entrenat el citat algorisme.

En alguns casos, però, els encerts en la classificació s'analitzen més profundament: quan la classificació té només dues possibles opcions (que anomenarem "positiu" i "negatiu"), es pot calcular per separat la capacitat d'encert o error en cada un dels casos. D'acord amb això, es defineixen les següents magnituds, suposant que ambdues opcions de classificació (o classes) són disjuntes:

- TP (*True positives*): aquells casos positius que han estat correctament classificats.
- TN (*True negatives*): aquells casos negatius que han estat correctament classificats.
- FP (*False positives*): aquells casos negatius que, erròniament, han estat

classificats com a positius.

- FN (*False negatives*): aquells casos positius que, erròniament, han estat classificats com a negatius.

Aquest estudi porta a la definició dels conceptes de sensitivitat i especificitat. La primera mesura la capacitat de classificar correctament els exemples positius, respecte tots els casos positius. La segona la capacitat de classificar correctament els exemples negatius, respecte tots els casos negatius:

$$sens = \frac{TP}{TP + FN} \qquad (3.1)$$
$$spec = \frac{TN}{TN + FP}$$

La representació d'aquestes magnituds en un sistema de coordenades duu al que es coneix com una corba operacional (*Receiver Operating Characteristic*, ROC, [58]), d'important utilitat si el problema respon a una de les següents situacions: d'una banda, existeix la possibilitat que la funció de cost no sigui homogènia. És a dir, que degut al significat de què és positiu i què és negatiu, cometre un error en un sentit sigui més costós que en un altre. Això és bastant habitual en l'entorn mèdic: si es comet un error en un diagnosi, la gravetat d'aquest fet per al pacient depèn del sentit de l'error. També és així en els problemes de detecció de fraus ([1]), on un frau no detectat és molt més greu que una falsa alarma.

D'altra banda, alguns algorismes de classificació basen la seva decisió en l'anàlisi d'una funció de pertinença a partir d'un cert valor llindar, que marca a quina classe pertany aquell cas. Dos algorismes no tenen el mateix comportament si prenen la mateixa decisió de classificació però ho fan amb iguals llindars i valors de la funció de pertinença molt diferents. Sempre es considerarà millor ("més robust", es dirà) aquell pel qual la diferència entre el valor de pertinença i el llindar és major.

Per ambdós casos, l'estudi de la bondat de l'algorisme a partir d'una corba ROC aporta avantatges sobre la utilització de la precisió o l'error de classificació. Aquestes corbes són habitualment convexes representades en l'espai (FT, TP), i passen pels punts $(FP, TP) = (0, 0)$ i $(FP, TP) = (1, 1)$: se'n mostra un exemple a la figura 3.2, extreta de l'article de Provost, Fawcett i Kohavi ([1]). Es demostra que l'àrea sota una corba ROC (coneguda per les inicials AUC, de *area under the curve*) és un bon indicador de la bondat de l'algorisme corresponent i, per tant, el valor d'aquesta magnitud també

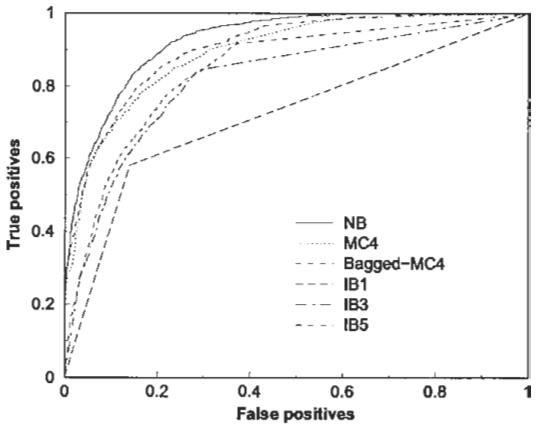

Figura 3.1: Exemple de corba ROC extret de l'article de Provost, Fawcett i Kohavi ([1]), on es pot veure la relació entre TP i FP per tot un conjunt d'algorismes classificadors. Tal i com s'ha mencionat, les corbes són convexes i passen pels punts $(0,0)$ i $(1,1)$.

pot ser considerat com una altra possibilitat en l'estudi comparatiu d'M algorismes.

En el cas que ens ocupa, l'AUC és d'un ús encara relativament reduït, malgrat les diverses opinions expressades des de la pròpia comunitat ([1], [35]) sobre els riscos que implica realitzar comparacions tenint en compte només la precisió o l'error de classificació. Alguns autors, com Demsar ([9]), opinen que la gran quantitat de problemes de prova disponibles per assajar els algorismes facilita l'ús amb suficient garanties de la precisió o l'error en la classificació. De fet, l'enfocament habitual és difícilment compatible amb distribucions de classes molt diferents a les que apareixen al problema de prova (és conegut i sovint referit el cas del problema *IRIS*, [12]), o bé amb aquells casos en què el cost de l'error de classificació no és igual per totes les classes. En aquests casos, és més recomanable utilitzar l'AUC com a mesura de bondat.

El desenvolupament que fa del problema aquest treball es basa en casos en què la precisió o l'error són les mesures amb què s'avalua el comportament de l'algorisme. No obstant això, si aquesta mesura es fes amb l'AUC, tot el que segueix tindria validesa, i faria falta tan sols adaptar algunes conclusions. En tot cas, és feina de futur valorar quins aspectes de l'anàlisi esdevenen més senzills tenint en compte tota la informació que s'inclou a les corbes ROC,

que és major que el simple valor d'AUC. En aquesta línia, també caldria estudiar la proposta de corbes de cost feta per Drumond ([59]).

3.3 Fonts i tipus d'error

En aquest apartat s'analitzarà l'origen i les causes dels errors que es poden produir en la tasca encomanada a un algorisme, identificant el marc teòric en què s'estudiarà el problema i separant aquelles fonts d'error que són exclusivament atribuïbles al propi algorisme.

En primer lloc, cal dir que es particularitzarà el plantejament per als casos d'algorismes classificadors, doncs són els que s'utilitzen en el nostre àmbit de treball. Les qüestions que es centraran sobre l'estimació de l'error, però, seran del tot generalitzables a un algorisme de qualsevol altre tipus, mentre tingui clarament definida una tasca a realitzar i un indicador que n'avaluï el compliment dels objectius encomanats.

Seguint aquest argument, es determinen a continuació el marc i la nomenclatura en què es desenvolupa un problema de classificació, tenint present que l'objectiu serà estimar la bondat d'un algorisme en la tasca classificadora que té encarregada.

3.3.1 Marc classificador

Considerem X el conjunt d'instàncies o elements d'una determinada població, sobre la qual es voldrà dur a terme un procés de classificació, i que està format pels elements $\{X_1, ..., X_N\}$. Cada una de les instàncies X_i és un element del conjunt X, que segueix una distribució D. A banda, considerem Y el conjunt de les classes possibles per als elements del conjunt X, format pels elements $\{Y_1, ...Y_M\}$.

A priori, es considera que existirà una funció objectiu f que relacionarà els elements dels dos conjunts, i que serà allò que es vol descobrir o aprendre:

$$f : X \rightarrow Y \tag{3.2}$$

En el present treball s'utilitzaran sovint casos d'aplicació mèdica: per exemple, el conjunt X podria estar format per totes les mamografies realitzades i el conjunt Y contindria dues classes, en funció de si aquestes mostren una alteració benigna o maligna. La funció f és la relació (existent però desco-

neguda) que *mapeja* la relació entre els atributs de les mamografies i la seva classificació en benigna o maligna.

Sobre aquests conjunts, es defineix un classificador C com una aplicació que fa correspondre a cada instància X_i de X un element Y_j de Y:

$$C : X \rightarrow Y \tag{3.3}$$

l'objectiu del qual serà aprendre f, és a dir, reproduir en la mesura del possible els seus resultats. La relació cal que sigui unívoca: a cada element del conjunt X li correspon un i només un element del conjunt Y. Si no és el cas, el classificador C no realitza la tasca encomanada de classificació com s'espera: és difícilment avaluable un classificador que ni tan sols pot atorgar una classe a un element.

Per tal d'avaluar la bondat del classificador, s'assaja el seu comportament sobre un conjunt S, que és una mostra del conjunt X, amb elements $\{S_1, ..., S_{N'}\}$ i $N' < N$. Els elements d'aquest conjunt S segueixen una distribució D_S que no té perquè coincidir amb D: dependrà de com S reprodueixi la distribució que segueixen els elements de X. El problema de prova sobre el qual s'assaja el comportament d'un algorisme classificador C és precisament aquest conjunt S, i el grau de bondat del classificador vindrà donat per la capacitat de reproduir el comportament de la funció objectiu f.

3.3.2 Estimació de l'error

Habitualment, la bondat d'aquest classificador $C(X)$ s'interpreta en funció de la seva capacitat de reproduir el comportament de $f(X)$, és a dir, de la seva precisió en *predir* la classe que correspon a cada element de X. Ara bé, a la pràctica no es coneix el conjunt X, que representa el conjunt de la població, sinó un conjunt S que representa una mostra d'aquesta població, i que permet estimar el valor de la precisió.

Les qüestions a plantejar seran com construir aquest estimador i, un cop fet, quina n'és la seva precisió: quin és el biaix i la desviació, tenint en compte que s'ha avaluat a partir d'un conjunt $S \subset X$.

L'error que es trobarà en un classificador C, avaluant-ho sobre el conjunt S, es pot separar en diverses parts, depenent del seu origen. Sovint s'anomena a aquesta magnitud error aparent (e_a), o simplement error, i és degut a tres principals fonts:

- e_{est}: error d'estimació, l'error produït pròpiament pel procés estadístic de càlcul de l'estimador de l'error. Aquest error tendeix a desaparèixer

quan $S \rightarrow X$, doncs en aquest cas la comprovació del comportament de C es faria sobre tots els elements sobre els que actua $f(X)$[1].

- e_r: error real, aquella fracció d'elements de X mal classificats per C, assajat sobre tota la població X. Aquest error és el valor que es busca estimar, i apareix degut principalment a dues raons:
  - e_{inh}: error mínim inherent o error intrínsec, fruit de la incapacitat de la representació de X de què es disposa (a partir dels atributs que caracteritzen cada element X_i) per aportar la informació necessària que permeti una correcta classificació. Diríem que els atributs no poden representar l'espai X de manera complerta. Aquest és el que es coneix com a error òptim de Bayes ([60]) i, si els atributs aportessin prou informació, hauria de ser zero.
  - e_{lan}: l'error propiciat per la limitació del *llenguatge* amb què es construeix C per a representar a l'aplicació f. Un classificador lineal, per exemple, utilitza una representació del coneixement diferent a un classificador basat en els veïns més propers (NN), i això forçosament ha de tenir un efecte sobre la seva precisió en ser aplicat a S.

D'acord amb aquestes definicions, es pot expressar l'error aparent obtingut d'avaluar un classificador C sobre el conjunt S com:

$$e_a = \frac{1}{N'} \sum_{X_i \in S} (1 - \delta(f(X_i), C(X_i)) \tag{3.4}$$

on l'operador δ és la coneguda delta de Kronecker, que val zero excepte si els seus arguments coincideixen (en aquest cas, si $f(X_i) = C(X_i)$). De manera similar, l'error real es pot expressar com:

$$e_r = \frac{1}{N} \sum_{X_i \in D} (1 - \delta(f(X_i), C(X_i)) \tag{3.5}$$

d'on s'observa que l'única diferència apareix en el conjunt sobre el qual s'avalua el classificador C.

Aquest darrer serà un valor constant, funció dels propis conjunts d'elements i de la representació escollida per construir C, mentre que e_a variarà

[1] De manera similar al que preveu el teorema fonamental del càlcul per la precisió d'un estimador qualsevol sobre un univers d'elements, [?]

en funció del conjunt S i del procediment utilitzat per a fer-ne la mesura. En general,

$$e_a = e_r + e_{est} = e_{inh} + e_{lan} + e_{est} \tag{3.6}$$

i l'objectiu serà aconseguir un estimador e_a que s'acosti el més possible a e_r (per tant, amb un biaix molt petit o, el que és el mateix, amb $e_{est} \to 0$), i que retorni valors, en cada càlcul realitzat, molt propers a la seva mitjana (variància de l'estimador reduïda). Aquestes dues qüestions són les que determinaran la bondat d'un estimador.

Abans d'acabar aquest apartat, es desitja introduir una variant en el plantejament, que ja s'ha utilitzat en alguns casos publicats ([5]). De vegades, com s'ha comentat quan s'ha fet esment a les corbes ROC, la decisió de classificar es pren a partir d'una funció del pertinença i un llindar.

Una opció que pot tenir sentit, especialment en aquells problemes on el cost d'un error sigui molt elevat (o el cost d'un dels dos possibles errors, FP o FN), és variar la manera d'efectuar la pròpia classificació, tot renunciant a la classificació en aquells casos més "dubtosos". Això provoca un descens immediat de l'error, tot i que a canvi d'augmentar els elements no classificats. Ja que s'està discutint sobre l'error, és interessant contemplar la possibilitat de crear una nova classe "no coneguda" si amb això es redueix apreciablement l'error en la classificació: es classifica menys, però en aquells casos en que es fa, l'error és menor.

Un exemple es pot trobar en el problema del darrer article esmentat, on s'estudia la clusterització a través de mapes auto-organitzats (SOM, [61]) com a eina per millorar l'etapa de recuperació (*retrieval*) en el raonament basat en casos (CBR, aquest procés es descriurà amb més amplitud a l'apartat 4.2), per uns problemes de càncer de mama.

L'objectiu de la variant del CBR presentada és introduir arguments probabilístics en l'etapa de recuperació de la memòria de casos del CBR, per millorar la capacitat de classificació de l'algorisme. En concret, es modifica el mecanisme per classificar els nous casos provinents del problema de prova, tot mesurant el seu grau de semblança a 1, 3 o 5 veïns més propers (1-NN, 3-NN, 5-NN), i generant una nova variant del SOMCBR, que anomenarem $SOMCBR - per$.

El CBR clàssic no mostra canvis importants en la precisió, com tampoc ho fa la variant del SOMCBR que mesura la pertinença per un sistema de votació respecte els veïns més propers ($SOMCBR - vot$)[2]. En canvi, els

[2]Els resultats obtinguts per tots els algorismes utilitzats en aquest problema es poden veure a la taula 9.18.

sistema introduït com a $SOMCBR-per$ redueix el nombre d'elements que es classifiquen (amb percentatges de no classificats que en casos extrems poden arribar al 80% dels elements), però redueix també molt l'error en el cas en què es procedeixi a la classificació.

A la taula 3.1 es poden veure els resultats de $SOMCBR-per$, per 5 problemes de prova de domini mèdic: CA i BI ([54], [49]), DD, M3 i MB ([55], [56]). Els tres darrers casos s'han separat en problemes de dues classes (amb el procediment que es mostrarà a l'apartat 4.4) per tenir un conjunt de problemes de prova majors, i a la vegada analitzar els resultats també des del punt de vista de l'especificitat i la sensitivitat (la qual cosa es pot fer només per problemes de dues classes).

Prob. de	**1-NN**		**3-NN**		**5-NN**	
prova	%Err.	%No-cl.	%Err.	%No-cl.	%Err.	%No-cl.
CA	24.0	0.0	5.7	43.8	2.8	60.6
BI	37.3	3.3	9.5	60.1	4.1	78.1
DD c1	18.4	0.0	7.3	24.6	4.8	35.3
DD c2	37.7	0.1	9.9	61.6	4.6	79.5
DD c3	36.1	1.6	12.4	52.8	4.9	71.6
DD c4	25.5	0.1	9.0	37.8	4.9	71.6
M3 c1	27.0	1.0	8.7	48.7	3.1	68.7
M3 c2	16.4	0.9	6.1	32.7	3.5	50.9
M3 c3	32.2	1.5	11.6	51.7	5.3	72.9
MB c1	20.0	0.5	5.6	41.7	3.0	61.9
MB c2	31.3	0.1	7.5	52.1	3.4	69.2
MB c3	26.7	0.2	8.2	42.0	3.4	58.3
MB c4	14.3	1.0	6.0	21.7	4.7	31.4
Mean (σ)	26.7 (7.9)	0.8 (0.9)	8.3 (2.2)	43.9 (12.4)	4.0 (0.9)	62.3 (15.2)

Taula 3.1: Resultats per a l'algorisme $SOMCBR-per$, assajat sobre 12 problemes de domini mèdic, per als quals s'ha modificat el número d'elements veïns que consideren a la fase de recuperació de la memòria de casos. Es mostra, per cada variació, el percentatge d'error en la classificació i el percentatge de casos que no es classifiquen.

Només el resultat expressat en mitjana (amb totes les informacions que es perd amb aquesta operació) ja mostren el gran guany en precisió que es fa a mesura que augmenta el número de veïns que es tenen en compte en la fase de recuperació, doncs l'error de classificació es redueix apreciablement. Aquest guany s'aconsegueix a força d'augmentar els casos no classificats, però en un cas d'aplicació mèdica com aquest, el cost d'un error serà possiblement bastant més gran que el d'un no classificat. La tendència observada és la mateixa si l'anàlisi es fa separant els errors per especificitat i sensitivitat.

Aquest exemple serveix per reforçar la proposta plantejada en aquest apartat, parlant de les fonts de l'error i la seva mesura: depenent del problema, pot ser desitjable destinar tots els esforços a reduir els errors, tot i que això impliqui deixar sense classificar un nombre elevat de casos. En el fons, aquesta elecció transforma el problema en un cas multivariant, en què hi ha dues mesures de bondat: l'error en la classificació, i el percentatge d'elements no classificats.

A la figura 3.2 s'observa la comparació dels resultats obtinguts per les tres variants del $SOMCBR-per$. En aquesta gràfica, cada punt representa el resultat obtingut per les tres variacions de l'algorisme sobre un problema de prova. En el cas 1-NN es reprodueixen resultats molt similar als d'un CBR, mentre que a mesura que augmenten els veïns més propers considerats (3-NN i 5-NN) disminueix l'error però augmenta el percentatge de no-classificats. El punt òptim seria el $(0, 0)$, és a dir, error nul amb tots els elements classificats. En aquest cas, es fa patent el principal repte d'un problema multivariant: l'establiment d'una funció de cost o objectiu per determinar el pes de cada variable en el concepte de bondat, i d'aquesta manera veure quina configuració de l'algorisme és l'òptima.

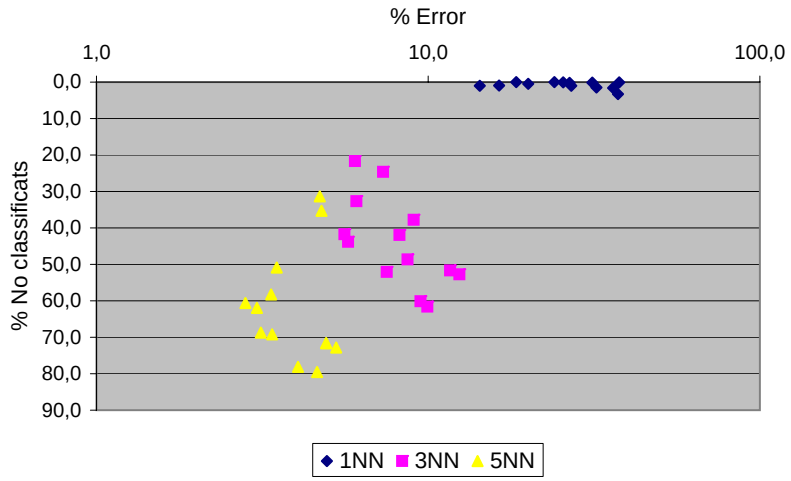

Figura 3.2: Representació dels resultats de l'algorisme $SOMCBR-per$ per a 1, 3 i 5 veïns més propers considerats a l'etapa de recuperació. S'hi representen l'error de classificació en l'eix d'abscisses (en un eix logarítmic per facilitar la visualització dels tres grups de dades), i el percentatge d'elements no classificats en l'eix d'ordenades. Cada punt representa el resultat obtingut per una configuració de l'algorisme (1-NN, 3-NN o 5-NN) sobre un problema de prova dels que apareixen a la taula 3.1.

3.4 Càlcul de la bondat

D'acord amb allò expressat a l'apartat precedent, l'objectiu és trobar un estimador de l'error d'un algorisme que tingui un biaix i una variància tan proper a 0 com sigui possible. Com es veurà, els mètodes desenvolupats topen sovint amb un compromís entre ambdues magnituds, pel qual reduir el valor d'un implica necessàriament augmentar el de l'altre. Els valors que prenguin i el tipus de problema seran els factors que determinin quina és l'opció òptima. Com és d'esperar, cap estimador serà sempre perfecte en el sentit donat de biaix i variància ([27], [34]).

La majoria dels mètodes desenvolupats resolen amb tècniques de *resampling* el problema principal: com realitzar l'aprenentatge del classificador i l'avaluació d'aquest aprenentatge comptant només amb una mostra S del conjunt X (és a dir, amb un problema de prova). La idea és extreure del conjunt S dos subconjunts amb els quals desenvolupar l'entrenament del classificador (S_{tr}, de *training*) i l'avaluació de l'error (S_{te}, de *test*).

Cal tenir en compte que, per diferents subconjunts $\{S_{tr}, S_{te}\}$, el classificador construït serà diferent i també ho serà l'avaluació: sovint es repeteixen diversos assajos per diferents conjunts $\{S_{tr}, S_{te}\}$, i se'n fa la mitjana per avaluar l'aprenentatge. D'acord amb la manera com es construeixen aquests conjunts $\{S_{tr}$ i $S_{te}\}$, i amb quantes iteracions es fan d'aquest procés, es poden classificar els estimadors de la següent manera:

- Aquells per els quals els mateixos elements serveixen per fer aprendre al classificador i per avaluar aquest aprenentatge.
- Aquells en què passa això mateix, però mai simultàniament: a cada iteració, els elements utilitzats per avaluar-ho son diferents dels utilitzats per fer-lo aprendre.
- Aquells en què els conjunts $\{S_{tr}, S_{te}\}$ es formen mostrejant el conjunt S amb la possibilitat de repetir elements (mostratge amb substitució).

En funció d'aquestes possibilitats es construeixen els estimadors que es descriuen tot seguit. En tots els casos, per comoditat en escriure les equacions, es treballarà sobre la precisió ($\%Acc$, de l'habitual *accuracy*) definida com el complementari de l'error. A banda, sovint també es construeixen tècniques híbrides ([28], [62]), però sempre a partir de les quatre grans opcions que aquí s'exposen. Per aquest motiu, no s'hi dedicarà una atenció especial.

3.4.1 Estimació per re-substitució

Una primera possibilitat consisteix en avaluar l'error del classificador C sobre el mateix conjunt S utilitzat per a entrenar-lo: és a dir, com en aquest cas $S_{tr} = S_{te} = S$, es pot descriure la precisió de l'algorisme com

$$\%Acc_{rs} = \frac{1}{N} \sum_{X_i \in S} \delta(f(X_i), C(X_i)) \tag{3.7}$$

on l'aprenentatge del classificador C s'ha realitzat amb el mateix conjunt S. L'error aparent que s'obté està esbiaixat de manera optimista respecte l'error real: s'observen millors resultats, perquè l'avaluació es fa directament sobre allò après, i no hi ha elements nous en el conjunt.

Aquest estimador funciona raonablement bé només en sistemes lineals i amb un volum de dades important ([63]), com sembla indicar la tendència asimptòtica a zero del biaix, ja prevista anteriorment per McLachlan ([64]). De fet, pel cas límit en què el classificador aprèn totes i cada una de les instàncies d'entrenament, aquest estimador retorna un 100% de precisió, que en cap cas significaria un comportament perfecte sobre un altre conjunt S' extret de X, i diferent de S.

3.4.2 Holdout

Per tal de superar bona part dels problemes presentats en el mètode anterior, s'acostuma a treballar amb una metodologia *holdout* o de subconjunts independents: donat el problema de prova S de què es disposa, es divideix aleatòriament en dos conjunts mútuament excloents (S_{tr} i S_{te}). El primer d'aquests conjunts és l'utilitzat per a l'aprenentatge del classificador, la bondat del qual s'estima sobre el segon conjunt:

$$\%Acc_{H} = \frac{1}{N} \sum_{X_i \in S_{te}} \delta(f(X_i), C(X_i)) \tag{3.8}$$

on C s'ha construit a partir de S_{tr}, i $S_{tr} \cup S_{te} = S$, amb $S_{tr} \cap S_{te} = \emptyset$.

La tria dels percentatges dels elements que seran membres de S_{tr} i S_{te} respon sovint a diverses motivacions, si bé s'acostuma a utilitzar dos terços dels elements de S per a l'entrenament i el terç restant per al test. Estadísticament, l'estimador és pessimista, en tant que no tots els elements de S són utilitzats per a l'entrenament.

La formació del conjunt S_{tr} i S_{te} respon a un compromís: l'estimador obté un valor més proper al real com més gran sigui el conjunt de test ([62]), però augmentar la mida de S_{te} implica reduir la de S_{tr}, amb la qual cosa l'aprenentatge de C serà pitjor. A més, el resultat dependrà molt de la tria dels elements que es faci per formar S_{tr} i S_{te}, per la qual cosa la variància de l'estimador és força elevada. De cara a reduir-la, sovint es repeteix l'experiment un cert nombre de vegades per particions aleatòries (*random subsampling*), mètode que molt fàcilment deriva cap als coneguts com a *cross-validation*, i que es presenten a continuació.

3.4.3 *k-fold cross-validation*

En la metodologia *k-fold cross-validation* (k-CV), el conjunt S original és aleatòriament dividit en k subconjunts disjunts d'aproximadament la mateixa mida (S_k). A partir d'aquí, s'entrena l'algorisme en $(k-1)$ dels subconjunts i s'avalua l'aprenentatge en el subconjunt restant, repetint el procés k vegades, tot variant el subconjunt $S_k = S_{te}$. La precisió del mètode es defineix com la mitjana per les k repeticions del quocient entre el número d'instàncies correctament classificades i el seu número total:

$$\%Acc_{k-CV} = \frac{1}{k} \sum_{runs} \left(\frac{1}{N} \sum_{X_i \in S_{te}} \delta(f(X_i), C(X_i)) \right) \tag{3.9}$$

on el classificador C està entrenat amb el conjunt S_{tr} format pels $(k-1)$ subconjunts S_k d'entrenament (diferents per cada una de les k iteracions), i S_{te} és el conjunt S_k no inclòs a S_{tr}.

Per valors de $k << N$ (número de *folds* molt menor que el número d'elements a S), el procés es pot repetir m vegades, i el resultat es calcula com la mitjana de les m mesures de $\%Acc_{k-CV}$ obtinguts: es parla llavors del *m-times k-fold cross-validation*. A nivell teòric es parla també de la possibilitat de realitzar una operació de *cross-validation* "complerta", entenent-ho com la repetició del procés per totes i cada una de les possibilitats d'escollir N/k instàncies en un conjunt de N. Aquesta possibilitat, però, implicaria repetir el procés un número de vegades igual a $\binom{N}{N/k}$, la qual cosa és del tot impossible[3].

[3]És interessant recordar la magnitud d'una quantitat d'aquest tipus. Com a nota curiosa, es pot reproduir el comentari irònic de [33]: *"One reviewer asked if we ever tried running complete cross-validation to show that it is better for our datasets. For the chess dataset, one would need to run the induction algorithm* $\binom{900}{90}$ *times. If every one of the*

Un cas extrem és aquell en el qual $k = N$, que dóna lloc al mètode conegut com a *leave-one-out* (LOO, [65]): es repeteix N vegades un procés en el qual el conjunt S_{tr} és format per tots els N elements de S menys un, que és utilitzat com a S_{te}. L'error final és la suma d'errors obtinguts per cada repetició, en tant que es mesura el número d'elements que no es poden classificar correctament, quan C ha estat construït amb tota la resta d'elements del conjunt S.

3.4.4 *Bootstrap*

Des d'un punt de vista similar al dels mètodes de Montecarlo, es poden intentar generar diferents conjunts d'entrenament (S_{tr}) i de test (S_{te}) a partir de la tria d'elements del conjunt S amb substitució. Així, per un conjunt S amb N elements, es pot crear tants conjunts d'N elements com es vulguin, escollint-los aleatòriament del conjunt S i tornant-los a aquest conjunt, mantenint una vegada i una altra la possibilitat de ser triats.

Com que per cada tria una instància determinada té $1/N$ de probabilitat de ser triada, la probabilitat de no ser escollida després de N tries esdevé

$$\left(1 - \frac{1}{N}\right)^N \tag{3.10}$$

que tendeix a l'invers del número e a mesura que N augmenta, doncs

$$\lim_{N\to\infty} \left(1 - \frac{1}{N}\right)^N = \frac{1}{e} \simeq 0.368 \tag{3.11}$$

A partir d'aquí, es pot calcular una mesura de la precisió del sistema classificador com:

$$\%Acc_{boot} = \frac{1}{m}\sum_{j=1}^{m}(0.632\%Acc_i + 0.368\%Acc_{rs}) \tag{3.12}$$

on m és el nombre de repeticions del procés, $\%Acc_i$ la mesura de precisió per la repetició i (entrenant el sistema amb el conjunt generat, i assajant-ne la precisió amb els elements de S que no han estat triats) i $\%Acc_{rs}$ la precisió segons el mètode de re-substitució sobre tot el conjunt d'elements, tal i com

10^{80} *atoms in this universe were replaced by a machine a million times faster than the Sparc 10 we used, and assuming C4.5 were optimized to be a million times faster, one would still need* 10^{25} *years, which we believe would miss some important deadlines."*

s'ha exposat abans. Efron ([28], [66]) va anomenar aquesta tècnica $e0$, i en va demostrar la semblança amb un *cross-validation* de dues particions del conjunt.

3.4.5 Estratificació

A l'inici d'aquest apartat s'ha comentat que l'objectiu és utilitzar un estimador que faci mínim l'error d'estimació, e_{est}. Tenint en compte que l'objectiu és avaluar la precisió de C a partir del coneixement que aporta el conjunt S, subconjunt del conjunt original X, ajudaria a assolir un millor resultat el fet que la distribució D_S fos el més semblant possible a D. Això és el que es busca quan es parla de l'estratificació del conjunt S.

El que pot passar és que, en el procés de divisió aleatòria del conjunt S per a formar S_{tr} i S_{te}, alguns dels subconjunts generats continguin una distribució molt poc equilibrada de les classes presents a S. Una millora en el càlcul es produeix amb l'estratificació, on es força que els subconjunts continguin aproximadament la mateixa proporció de cada classe que en el conjunt S original. En l'apartat final de resultats, es veurà com introduir aquesta pràctica en els estimadors duu a una apreciable millora, tant en biaix com en variància ([67], [68]).

3.4.6 Antecedents i resultats

Diversos estudis, la majoria d'ells estrictament experimentals, s'han dedicat en aquests darrers anys a comparar els estimadors exposats. Malgrat la majoria ho han fet sobre conjunts de dades molt petits, se'n poden extreure algunes conclusions generalitzables, orientades a respondre la qüestió cabdal d'aquest capítol: de quina manera escollir un estimador per al càlcul de l'error o la precisió de l'algorisme.

Al llarg de la dècada dels vuitanta i els primers noranta, equips com els encapçalats per Efron ([28]), Breiman ([29]), Jain ([30]) i Weiss ([31]), varen desenvolupar diferents treballs per comparar el biaix i la variància dels estimadors que anaven desenvolupant. Tots els experiments van ser realitzats sobre problemes de prova construïts *ad-hoc* als objectius del treball, o bé sobre problemes amb un número d'elements baix. Això va provocar que les conclusions fossin només parcials, i fins i tot en alguns casos contradictòries ([12]).

A partir d'aquests primers resultats, Kohavi ([32],[33]) i Martin ([12]) van

realitzar proves molt més extenses i tenint en compte elements nous en la representació del resultat o el rang d'aplicació dels classificadors (per exemple, en un problema que sigui linealment separable no es poden comparar els resultats que s'obtinguin amb un classificador lineal respecte els obtinguts amb un classificador basat en els veïns més propers).

D'aquests treballs se'n poden extreure un seguit de conclusions, que serviran de guia per a la tria de l'estimador òptim en cada problema:

- Tant l'estimació per re-substitució com per *holdout* són mètodes que provoquen elevats valors del biaix i la variància, fet que en general els descarta. La iteració del segon millora el resultat, però segueixen sent preferibles altres opcions.
- Els mètodes basats en el *cross-validation* (tant el k-CV com el LOO) són els que mostren un biaix menor en la gran majoria de casos (un biaix que sempre és pessimista). La variància és major, però en el cas del k-CV es pot reduir iterant el procés (*m-times k-fold cross-validation*), i en tots els casos és computacionalment menys costós que el LOO. A més, Kohavi demostra que un valor de k proper a 10 optimitza el compromís entre biaix i variància.
- El *bootstrap* mostra en general una variància menor que el k-CV, però el problema és que el biaix no es pot controlat: diversos autors consultats troben casos en què el comportament a aquest nivell és excel·lent (podem parlar fins i tot d'un estimador no esbiaixat), i d'altres en què és molt dolent. Això desaconsella el seu ús, especialment en aquells problemes on es vulguin comparar comportament d'algorismes classificadors basats en lògiques diferents.
- En tots els casos, l'estratificació millora els resultats del biaix, i també de la variància.

Tots aquests elements porten a afirmar que una bona recomanació és la utilització del *m-times k-fold cross-validation* amb estratificació ([10]), com es farà en la majoria dels problemes que s'estudiaran al llarg d'aquest treball: l'únic cas en què pot ser recomanable una alternativa és si la mida del conjunt S és molt gran. En aquest cas, es pot utilitzar el *holdout*, a no ser que el seu cost computacional sigui massa elevat ([31]).

Una variant del k-CV és el conegut com a 5x2-CV, en què es fan 5 iteracions d'un *cross-validation* amb 2 *folds*. La variació és mínima: Dietterich ([13]), per exemple, a partir de la prova sobre classificadors $C4.5$ i $k-NN$

i amb el t-test com a test d'inferència, mostra com el seu funcionament és molt similar. Malgrat el 5x2-CV redueix la probabilitat de cometre un error "Tipus I" i el k-CV ho fa amb l'error de "Tipus II", les diferències són molt petites, i la recomanació feta segueix essent vàlida.

La clau de volta de tot plegat l'acaben donant Martin i Hirschberg: l'important és mantenir un criteri homogeni en la metodologia per la qual es troben els resultats de la bondat de cada un dels algorismes que es vulguin comparar ([12]). Les petites diferències que poden haver-hi entre el valor trobat i l'error real, si es segueixen les recomanacions anteriors i no es canvia de criteri enmig d'un problema, seran molt menors que les diferències entre l'error real que existiran entre els diferents algorismes que es compararan, molt especialment si entre aquests apareixen diferències significatives.

3.4.7 Altres aspectes a comentar

Complementàriament a tot l'anàlisi fet fins ara, es comentaran tres qüestions més, que poden tenir una certa importància en algun cas concret. La primera fa referència a l'anàlisi de la variància de l'estimador, que ha estat en sí mateix un tema d'estudi, especialment els darrers anys. Per exemple, Bengio i Grandvalet ([40]) han estudiat abastament l'estimador de la variància del *k-fols cross-validation* com a tècnica per estimar de la precisió.

S'ha dit en l'apartat anterior que el k-CV és un estimador de l'error amb un biaix molt petit, i que per reduir-ne la variància es poden iterar els mètodes m vegades. Doncs bé, aquests autors han arribat a demostrar la gran dificultat existent per al càlcul d'aquesta variància, fins al punt de trobar que no existeix un estimador no esbiaixat de la variància del k-CV. Davant de conclusions com aquestes, encara prenen més importància els comentaris del final de l'apartat anterior, sobre quines són les recomanacions essencials a seguir.

Una altra qüestió que també cal tenir en compte és el que es coneix com a sobre-aprenantatge o sobre-especialització (*overfitting*, [69]), definit per Quinlan ([70]) com l'augment de la complexitat del classificador amb l'única motivació d'adaptar-se a un cas especial, representat per un element concrets de S. Si el sistema "aprèn en excés", un algorisme classificador C pot mostrar un error molt menor respecte el conjunt $S \in X$ que l'ha generat, però un error molt gran en un altre $S' \in X$ que contingui elements de X que no estaven en S.

Pel cas del k-CV, Burman ([65]) demostra que el biaix és de l'ordre de

$$\frac{O(p)}{(k-a)N} \tag{3.13}$$

on p és el nombre de paràmetres que ajusten el comportament del classificador i N el nombre d'elements de S. Un classificador sobre-especialitzat ($p \sim N$) té un biaix molt elevat i, en general, aquest creix amb p. en alguns classificadors concrets, de vegades es diu que l'algorisme està saturat, quan el número de paràmetres a ajustar s'acosta al dels casos per entrenar-lo.

Com sempre, la solució rau en mantenir l'equilibri, i rebutjar un augment desmesurat en els paràmetres que calgui resoldre per calibrar l'algorisme classificador. En aquest tema s'està sempre sotmès a la llei de conservació de la generalització enunciada per Schaffer ([27]), coneguda també per principi del "*no-free-lunch*" ([71]). El propi Schaffer o altres com Breiman i Spector ([72]) troben en el *cross-validation* una metodologia que es situa en un bon compromís donats tots aquests problemes.

Finalment, en el cas de conjunts S amb un número molt elevat d'elements, una bona opció pot ser el concepte de corba d'aprenentatge utilitzat per Cortes i altres ([73]), i que no ha tingut gaires seguidors posteriorment. El concepte és el següent: si el número d'elements de S és molt elevat, i el procés d'entrenament o el de test és molt costós, podria tenir sentit fer aquest processos per subconjunts de S cada vegada amb més elements, i estudiar l'evolució de la precisió: és d'esperar que aquesta augmenti, i potser es pugui extrapolar el límit asimptòtic al qual tendeixi aquesta "corba d'aprenentatge".

Kohavi ([32],[33]) ha realitzat la prova tot augmentant la mida de S_{tr} i mantenint fixa la mida de S_{te}. En alguns casos, el comportament asimptòtic vers la precisió es mostreja per valors petits de S_{tr}, i amb un model no paramètric i una aproximació utilitzant el mètode de Lavenberg-Marquart ([74]), s'aconsegueix una mesura de la precisió amb un biaix molt petit.

Probablement, la proposta no ha tingut posteriorment molt de suport degut al cost que implica aquest càlcul, i al fet que difícilment la mida de S és prou gran i el classificador C prou lent com perquè surti a compte estudiar la corba d'aprenentatge, si es considera en comparació amb l'aplicació d'un k-CV per a tot el conjunt S.

3.5 Resum

Com una de les etapes prèvies a la comparació del comportament d'un conjunt d'M algorismes, en aquest capítol s'han establert les bases que permeten calcular els indicadors que determinen la seva bondat i, a més, conèixer quina és la fiabilitat dels mètodes amb els quals es calculen.

En primer lloc, s'ha definit què s'entén per bondat d'un algorisme, centrant-se en els casos dels classificadors, i s'han comentat les diferents mesures que en poden donar una idea: error, precisió, AUC, etc. A continuació, s'ha presentat la nomenclatura a utilitzar en l'estimació d'aquestes quantitats, establint el marc teòric en què es situa un problema classificador. En aquesta línia, s'han definit les fonts que provoquen error en el comportament d'un algorisme, quines són aquelles que es pretenen avaluar, i s'ha mostrat un cas pràctic en el qual la reducció de l'error de classificació és possible si també es redueixen els elements que es classifiquen.

A continuació, i remarcant el fet que el càlcul de la bondat es realitzarà a través d'un estimador estadístic, s'han exposat les diferents tècniques que permeten superar el problema bàsic amb què cal enfrontar-se en mesurar aquesta bondat: com obtenir un valor si només es disposa d'un conjunt S d'elements (problema de prova), i aquest ha de servir per a la construcció del classificador (entrenament) i la pròpia avaluació (test).

Tots els mètodes habituals són exposats i se'n remarquen els problemes des del punt de vista de les dues magnituds que es volen minimitzar: el biaix i la variància de l'estimador. Aquest apartat finalitza amb un recull dels resultats que es poden trobar en treballs fets anteriorment, i amb un seguit de conclusions que se'n desprenen i que funcionen a mode de guia per triar l'estimador en cada problema.

Finalment, es discuteixen tres aspectes més d'aquestes metodologies (molt especialment en el cas del k-CV, que acaba sent la millor opció en la majoria dels casos), pel seu interès concret en determinats problemes.

Capítol 4

Estudi de les propietats inherents del problema

"Però què s'entén quan es fa servir la paraula complexitat en l'àmbit de les ciències?
En realitat, ni els propis científics es troben d'acord sobre una definició única."
http://einsteinalaplatja.blogspot.com

Per analitzar el comportament d'un conjunt d'algorismes, cal assajar-los sobre una col·lecció de problemes de prova, a partir d'algunes de les tècniques exposades al capítol anterior. A partir d'aquí, aquest capítol estudia quin és l'efecte d'analitzar aquest comportament sense una acurada anàlisi de les propietats inherents d'aquests problemes. Es vol fer patent l'error d'obviar aquesta anàlisi, i la manera de procedir davant aquesta situació, tot relacionant-ho amb l'estudi de la complexitat d'un problema.

Utilitzant l'aplicació de diferents variants del raonament basat en casos (que s'introduiran convenientment) sobre dues col·leccions de problemes de prova (utilitzades a [75] i [22]), es mostraran diversos exemples que indicaran l'efecte que les propietats inherents als problemes utilitzats tenen en el resultat obtingut. A partir d'aquests exemples, es mostrarà com les mètriques de complexitat són una bona eina per a l'estudi previ dels problemes de prova, i s'introduirà una nova mesura que farà possible la definició d'un mapa amb diferents regions de complexitat. Aquestes regions determinaran la metodologia d'aplicació dels test estadístics per a l'anàlisi dels resultats, que seran desenvolupats als capítols 6 i 7.

4.1 Plantejament

Tal i com s'ha introduït en capítols anteriors, una situació habitual és l'estudi del comportament de M algorismes, a partir del resultat obtingut sobre N problemes de prova. Sovint, la metodologia és ben simple: s'obtenen els valors de bondat per cada un dels algorismes sobre cada problema de prova a partir de l'estimador escollit (tal i com s'ha definit a la primera part d'aquest treball, obtenint $X_{i,j}$ com a mesura de precisió, error, etc.), i es comparen amb les tècniques que es mostraran a partir del capítol següent.

Ara bé, en molts pocs casos es fa una anàlisi relativament profunda de les característiques dels problemes de prova. Els test estadístics que es presentaran més endavant parlaran de la necessitat que els N problemes de prova hagin estat escollits a l'atzar dins l'univers de tots els problemes de prova existents, però no determinarà més condicions. Alguns autors, en canvi, analitzen unes tipologies de problemes particulars, ja sigui a través de les pròpies característiques de les dades (per exemple en els problemes no amb distribució de classes poc equilibrada, [76]) o dels àmbits d'on provenen els problemes (per exemple dominis mèdics, [77]).

No obstant això, no es troba a la bibliografia existent una costum estesa d'analitzar els problemes de prova prèviament a l'aplicació dels algorismes, la qual cosa pot dur a conclusions errònies: de fet, si no fos així no tindrien sentit els referits estudis sobre col·leccions particulars de problemes. Una primera hipòtesi a comprovar, per tant, afirma que no és correcte l'estudi del comportament d'un conjunt d'algorismes a partir de l'aplicació sobre una col·lecció de problemes de prova sense analitzar prèviament les característiques inherents d'aquests problemes.

De quina manera cal estudiar prèviament aquestes dades o, dit d'una altra manera, l'anàlisi de quines característiques ens portarà la informació necessària per als nostres objectius? Una segona hipòtesi a comprovar expressa que les conegudes com a mètriques de complexitat ([78]) aportaran el coneixement suficient com per concloure correctament sobre el que es defineix com a domini de competència d'un algorisme ([79]): aquell conjunt de problemes, caracteritzat per un conjunt de propietats, pels quals aquest algorisme tindrà un bon comportament (definit això de la manera que en cada cas correspongui).

De la comprovació d'aquest parell d'hipòtesis que es proposen s'obtindrà una metodologia per a l'estudi del domini de competència d'un algorisme a partir de magnituds inherents al problema de prova i, per tant, avaluables prèviament a l'aplicació del citat algorisme. Especialment en aquells casos

en què l'aplicació de l'algorisme implica un temps de càlcul important, serà interessant poder obtenir alguna mesura *a priori* que indiqui si l'algorisme en qüestió retornarà un resultat d'acord amb els nostres interessos.

El cas particular que es presenta per comprovar aquestes hipòtesis es basa en una variant del raonament basat en casos (CBR) presentada a [21] i coneguda com SOMCBR. Aquesta variant, en les diferents opcions que s'enumeraran més endavant, implica un entrenament costós en termes de temps de càlcul: evitar l'aplicació de l'algorisme si ja es coneix *a priori* que el resultat que s'obtindrà no serà satisfactori, degut a les propietats inherents del problema de prova, pot estalviar càlculs llargs i que no aportin resultats d'interès. Dit d'una altra manera, aquesta sistemàtica permet una orientació prèvia cap a l'algorisme que, presumptament, ens donarà els millors resultats finals.

Ens els apartats que segueixen a continuació, es mostraran els resultats obtinguts fins arribar a la conclusió que les hipòtesis plantejades són certes. En primer lloc, a l'apartat 4.2 s'exposaran breument les variants del CBR utilitzades com a algorismes i els problemes de prova amb els quals es treballarà. A continuació, l'apartat 4.3 mostrarà els primers resultats que indiquen com existeixen diferències de comportament dels algorismes en funció dels problemes de prova als quals s'apliquen, en alguns casos molt extremes.

Un cop exposat aquest fet, es discutiran diverses mesures que permetin valorar prèviament la viabilitat de l'algorisme utilitzat, concloent la utilitat de les mètriques de complexitat (4.4). A partir d'aquí, a l'apartat 4.5 es defineixen i calculen les principals mesures de complexitat i els resultats obtinguts per tots els problemes estudiats. Finalment, a 4.6 es resoldrà la qüestió que apareix en estudiar aquests resultats: quina o quines mètriques de complexitat permeten la separació en grups dels problemes de prova (el que es definirà com a regions de complexitat), i quina és la relació amb els resultats obtinguts de l'aplicació dels algorismes estudiats.

El capítol acaba amb les principals conclusions (apartat 4.8), ampliables per a una metodologia general d'estudi d'M algorismes a partir dels resultats obtinguts sobre N problemes de prova: l'anàlisi previ de les propietats inherents als problemes de prova utilitzats pot evitar errors en l'estudi de la bondat dels algorismes, i les mètriques de complexitat són magnituds adequades per a realitzar aquest estudi.

4.2 Algorismes i problemes de prova

El raonament basat en casos (CBR, [80]) és una metodologia basada en resoldre nous problemes a partir de l'adaptació de solucions conegudes de problemes similars, resolts amb anterioritat. La seva formalització inclou quatre etapes ben definides, referides habitualment pels seus noms en anglès: *retrieval* (recuperació), *reuse* (adaptació), *revise* (revisió) i *retain* (emmagatzematge).

A la figura 4.2 ([50]) es mostra com el cicle d'aquest mètode gira al voltant d'una memòria de casos (MC), on s'emmagatzema el coneixement del sistema, habitualment en forma dels distints casos coneguts (entenent per cas la representació d'un problema amb la seva respectiva solució o classificació, a partir d'un conjunt d'atributs). Un cop preparat el nou cas amb aquells atributs que permeten la seva comparació amb el coneixement emmagatzemat a MC, s'inicien les fases mencionades:

- *Retrieval* (Recuperació): a partir del nou cas, es busquen a la memòria de casos tots aquells que compleixen uns certs criteris de similitud (per la qual cosa caldrà definir la funció de similitud entre casos). En el CBR la comparació es fa amb tots els elements de la MC, mentre que en les variants proposades a [75] es trien els elements amb els quals es fa aquesta comparació.

- *Reuse* (Adaptació): amb els casos de la MC més similars al nou cas, es proposa la seva adaptació com a solució, ja sigui a partir directament del cas de major similitud, d'una combinació lineal entre els casos més semblants, d'algun tipus de tria per votació, etc.

- *Revise* (Revisió): un cop es té la solució proposada, es revisa la seva validesa, amb un expert o a partir de certes regles de validació. Si és correcte, s'extreu la solució.

- *Retain* (Emmagatzematge): en alguns casos, i d'acord amb la política d'aprenentatge, es guarda el nou cas dins la MC. Aquí és on es produeix l'aprenentatge pròpiament dit, i és una fase altament crítica: una mala política pot arribar a tornar la MC inestable, i fer molt difícil la repetició del procés per nous casos.

D'aquesta breu explicació del CBR cal retenir una informació que serà cabdal en els apartats següents: si no es fa res d'especial, la fase de recuperació implica la comparació del nou cas amb tots i cadascun dels casos de

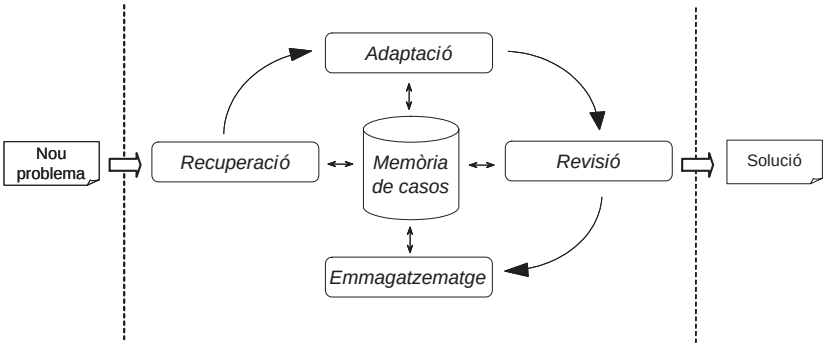

Figura 4.1: Esquema del cicle del CBR.

la MC. L'objectiu de les variacions que es proposaran sobre el CBR és, precisament, establir criteris per a reduir el número de comparacions a realitzar en aquesta fase. A banda, sovint a la MC hi apareix informació redundant o amb soroll, que influeix negativament sobre la qualitat de la recuperació. La reducció del nombre de comparacions pot ajudar a reduir la magnitud d'aquest fenomen.

Davant aquests fets, el SOMCBR [21] es proposa com una nova estratègia per a l'organització de la memòria de casos introduint-hi un procés de clústeterització basat en els mapes auto-organitzatius de Kohonen (Self-Organized Maps (SOM, [81]). Aquest pas intermig té l'objectiu d'arribar a un compromís acceptable entre els dos requeriments necessaris que s'han de complir: una resolució del problema amb un cost computacional controlable, i una qualitat acceptable del resultat (habitualment expressat en forma d'un percentatge d'encert en el procés de classificació).

A partir dels resultats obtinguts a [21], es van proposar algunes alternatives a l'etapa de recuperació, essent d'especial interès les presentades a [75]. En aquesta contribució, es proposa tenir en compte el paper d'aquells clústers que no ocupen el primer lloc en la llista dels més similars a la instància que s'intenta resoldre, a partir de diverses possibilitats: prendre en consideració tots els elements dels clústers amb una similitud major que un cert llindar (*elements from the best neighbours*, EBN), fer-ho només amb alguns elements d'aquests clústers (*a part of the elements from the best neighbours*, PEBN), o considerar elements de tots els clústers, en funció del seu grau de similitud (*an opportunity for all the neighbours*, OAN). Aquestes propostes busquen millorar els resultats obtinguts respecte el que implica el SOMCBR habitual, que utilitza només elements del clúster amb major similitud (*only the*

best model, OBM); tot plegat, intentant mantenir un número d'operacions de recuperació similars, en ordre de magnitud.

Totes aquestes propostes s'han aplicat sobre un conjunt de 14 problemes de prova, explicitats a la taula 4.1. Per tal d'obtenir resultats el més generals possible, s'han utilitzat dades que provenen de problemes sintètics i de problemes reals, alguns d'ells originaris del repositori UCI ([3]) com són *Iris, Heart-Statlog, Glass, Breast Cancer Wisconsin, Vehicle, Segment, Ionosphere* i *Sonar*. *Tao* és un problema de prova que representa la classificació dels punts en un cercle que representen el conegut símbol del *ying-yang*. La resta provenen de dominis mèdics, relacionats amb el diagnosi de càncer de mama: *Biopsy* ([54]) i μCa ([49]) estan construïts sobre mamografies prèviament classificades per l'Hospital Trueta de Girona, mentre que *DDSM* ([82],[55]) i *MIAS* ([83],[56]) provenen de bases de dades públiques. La diferència entre *MIAS-Birads* ([84]) i *MIAS-3C* prové només de la classificació realitzada.

4.3 La correlació entre el temps de càlcul i la precisió

Un cop presentats els algorismes i problemes de prova amb què es treballarà, es mostren a continuació alguns dels resultats obtinguts, d'on s'observaran comportaments molt diferents d'un mateix algorisme en funció del problema de prova sobre el qual s'ha assajat.

En primer lloc, cal dir que les diferents variants del SOMCBR comentades a l'apartat anterior s'han implantat per tres mides diferents de mapa de Kohonen: 3x3, 4x4 i 5x5. Els resultats obtinguts es poden consultar a l'article original de A. Fornells i altres ([75]), i s'utilitzaran parcialment per a l'assaig de les tècniques estadístiques que es presentaran més endavant.

L'interessant per a la nostra comesa és mostrar el distint comportament observat en alguns problemes de prova, que a més puguin explicitar els casos extrems de comportament. A les taules 4.2 i 4.3 es mostren els resultats obtinguts pels problemes *Iris* (IR) i *MIAS-Birads* (MB), pel que fa a la precisió obtinguda en el problema de classificació (*accuracy rate*, $\%AR$) i el número d'operacions en la fase de recuperació de la memòria de casos ($\#Op$). Cada columna de les taules mostren els resultats per una variant possible del CBR.

L'objectiu de les propostes presentades era reduir el temps de càlcul d'una manera significativa, tot mantenint la mateixa precisió que s'obté utilitzant un sistema CBR sense clústerització de la memòria de casos. Més endavant es discutirà si això s'aconsegueix i de quina manera, perquè ara l'objectiu és

Sigles	Prob. prova	Atributs	Classes	Distribució per classes
IR	Iris	5	3	iris-setosa (50), iris-versicola (50), iris-virginica (50)
HS	Heart-Statlog	14	2	absent (150), present (120)
GL	Glass	10	5	buildfloat (70), buildnonfloat (76), vehicle (17), containers (13), tableware (9), headlamps (29)
WS	Wisconsin	10	2	benign (458), malign (241)
VE	Vehicle	19	4	0 (212), 1 (217), 2 (218), 3(199)
SE	Segment	20	7	brickface (330), sky (330), foliage (330), cement (330), window (330), path (330), grass (330)
IO	Ionosphere	35	2	b (126), g (225)
SO	Sonar	61	2	rock (97), mine (111)
TA	Tao	3	2	black (944), white (944)
CA	μCa	22	2	benign (121), malign (95)
BI	Biopsy	25	2	0 (530), 1 (497)
DD	DDSM	143	4	b1(61), b2(185), b3(157), b4(98)
MB	MIAS-Birads	153	4	b1(128), b2(78), b3(70), b4(44)
M3	MIAS-3C	153	3	fatty(106), dense(112), glandular(104)

Taula 4.1: Descripció dels problemes de prova utilitzats per a l'assaig de les diferents propostes de SOMCBR, provinents de diversos repositoris. El primer grup són problemes del repositori UCI, mentre que el segon grup són problemes amb què treballa habitualment el Grup de Recerca en Sistemes Intel·ligents. De cada un d'ells s'indica l'habitual abreviació, el nom complert, el número d'atributs de cada instància (contant-hi la classe a la qual pertanyen), el número de classes, i la distribució de les instàncies per cada classe.

estudiar els resultats obtinguts en funció del número d'operacions efectuades en la fase de recuperació i, sobretot, si existeix alguna relació entre $\%AR$ i $\#Op$, depenent del problema de prova.

L'escenari desitjable seria aquell en què una disminució del número de comparacions ($\#Op$, que implicaria una reducció del temps de càlcul) no impliqui necessàriament una reducció en la precisió del resultat ($\%AR$). A la figura 4.2 es pot observar la representació gràfica dels valors de $\%AR$ respecte $\#Op$ per a dos problemes seleccionats, que mostren comportaments extrems.

En el cas del *Iris* es pot veure com la precisió del resultat no depèn del nombre de comparacions efectuades ni, per tant, del nombre d'elements de la memòria de casos que s'utilitzen en la fase de recuperació. Els valors de

	OBM		**EBN (ϑ=0.8)**		**PEBN (ϑ=0.8)**	
	%AR(σ)	**#Op**	**%AR(σ)**	**#Op**	**%AR(σ)**	**#Op**
SOMCBR 3×3						
IR	93.3 (5.9)	31	89.3 (5.3)	117	87.3 (6.9)	28
MB	59.1 (6.4)	57	69.1 (12.1)	288	57.2 (11.5)	44
SOMCBR 4×4						
IR	95.3 (4.2)	95	76.0 (18)	106	73.3 (22.3)	30
MB	64.7 (14.5)	165	69.1 (12.1)	288	57.8 (5.6)	41
SOMCBR 5×5						
IR	89.3 (5.3)	38	90.7 (4.9)	126	91.3 (4.2)	46
MB	69.1 (12.1)	288	69.1 (12.1)	288	68.4 (9.5)	288

Taula 4.2: Resultats obtinguts sobre els problemes *Iris* i *MIAS-Birads*, per tres de les configuracions assajades. Es mostra el percentatge mitjà d'encerts (%AR), la corresponent desviació estàndard (σ), i el nombre mitjà d'operacions a realitzar (#Op) necessàries per a recuperar un cas amb cada una de les configuracions exposades, utilitzant mapes de mida 3×3, 4×4 i 5×5 en el SOM. Els valors que es mostren per EBN i PEBN corresponen al llindar utilitzat internament per determinar quins clústers es tenen en compte, i de quina manera es recuperen els casos.

	PEBN (ϑ=0.5)		**OAN (x_0=0.8)**		**OAN (x_0=0.5)**	
	%AR(σ)	**#Op**	**%AR(σ)**	**#Op**	**%AR(σ)**	**#Op**
SOMCBR 3×3						
IR	93.3 (5.9)	26	84.0 (9.5)	94	85.3 (8.3)	124
MB	59.4 (7.7)	45	66.6 (10.8)	247	67.5 (11.5)	274
SOMCBR 4×4						
IR	92.7 (6.3)	40	83.3 (9.1)	94	83.3 (9.1)	124
MB	57.8 (5.6)	41	69.4 (8.6)	246	66.3 (10.2)	275
SOMCBR 5×5						
IR	95.3 (5.2)	42	84.7 (9.9)	94	84.0 (9.1)	124
MB	57.8 (7.3)	50	66.3 (11.2)	247	66.6 (11.7)	275

Taula 4.3: Resultats obtinguts sobre els problemes *Iris* i *MIAS-Birads*, per tres de les configuracions assajades. Es mostra el percentatge mitjà d'encerts (%AR), la corresponent desviació estàndard (σ), i el nombre mitjà d'operacions a realitzar (#Op) necessàries per a recuperar un cas amb cada una de les configuracions exposades, utilitzant mapes de mida 3×3, 4×4 i 5×5 en el SOM. Els valors que es mostren per PEBN i OAN corresponen al llindar utilitzat internament per determinar quins clústers es tenen en compte, i de quina manera es recuperen els casos.

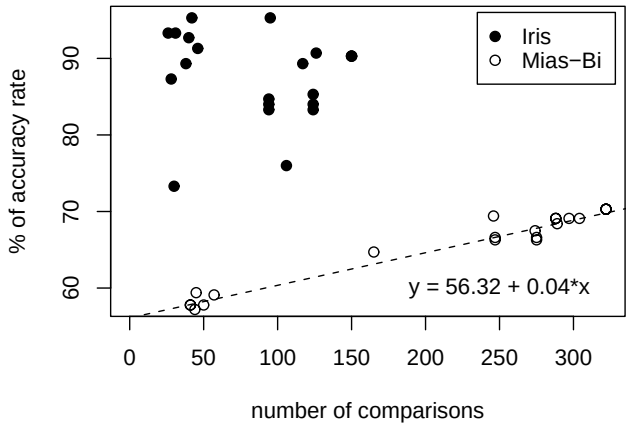

Figura 4.2: La figura mostra la relació entre la precisió del sistema i el nombre de comparacions en la fase de recuperació, pel cas dels problemes de prova *MIAS-Birads* i *Iris*. Cada punt de la gràfica representa un dels casos presentats a les taulse 4.2 i 4.3, és a dir, una estratègia de clusterització de la MC. Es pot observar com el primer d'ells mostra un comportament clarament lineal (es mostra també la recta de regressió lineal ajustada), mentre que el segon no mostra cap relació entre ambdós variables.

$\%AR$ varien de manera no determinada respecte els valors de $\#Op$. En canvi, en el cas del *MIAS-Birads* el resultat obtingut és clarament proporcional al nombre d'elements de la memòria de casos utilitzats. La relació és fortament lineal, amb un coeficient de correlació lineal $\rho = 0.98$, mentre que en el cas del *Iris* el coeficient val $\rho = 0.11$, coherent amb la no linealitat observada a la pròpia gràfica.

La conclusió que se'n pot extreure és la següent: en alguns casos (aquells amb ρ proper a 0), no hi ha cap relació de proporcionalitat entre $\%AR$ i $\#Op$. Per tant, serà possible mantenir la precisió obtinguda pel CBR a partir d'una estratègia que redueixi el nombre d'operacions a realitzar en la fase de recuperació. En canvi, en d'altres casos (aquells amb ρ proper a 1), la relació entre ambdós variables serà fortament lineal, i per tant no serà possible mantenir la precisió reduint el nombre d'operacions a realitzar.

El conjunt de valors de ρ obtinguts per als problemes de prova referits a la taula 4.1 es poden veure a la taula 4.4, on es mostren a més els millors resultats (en termes de precisió) per cada una de les estratègies proposades. Els problemes de prova estudiats es mostren separats en dos grups, en funció de si el valor d'aquest coeficient de correlació lineal és menor o major que

0.5.

	ρ	CBR		OBM		EBN		PEBN		OAN	
		%AR	#Op	%AR	#Op	%AR	#Op	%AR	#Op	%AR	#Op
TA	0.02	95.4	1670	94.7	544	86.2	926	93.6	201	53.1	994
IR	0.11	95.3	135	95.3	95	90.7	126	95.3	42	85.3	124
HS	0.15	74.1	243	76.3	54	74.1	243	78.1	50	78.5	243
WS	0.34	96.1	629	96.8	127	95.6	569	97.1	62	85.7	460
CA	0.39	62.5	195	67.2	140	63.9	195	63.9	34	66.7	190
DD	0.45	46.5	451	43.9	70	44.3	451	46.9	52	46.9	428
BI	0.49	83.2	925	83.2	924	83.2	925	82.4	925	82.2	791
IO	0.85	86.9	315	88.1	170	86.9	315	81.8	48	87.5	270
M3	0.91	70.1	288	65.4	45	70.2	288	64.9	50	72.1	276
GL	0.93	66.4	193	65.9	121	66.4	193	58.8	49	68.2	190
SO	0.94	87.0	187	84.1	84	87.1	187	73.1	48	87.9	161
VE	0.95	69.1	846	62.8	357	69.5	761	61.2	97	70.3	746
SE	0.96	97.3	2079	92.4	350	97.3	2079	90.6	244	96.1	1776
MB	0.98	69.1	288	69.1	288	69.1	288	68.4	288	69.4	246

Taula 4.4: Resultats per cada un dels problemes de prova, seleccionant la millor configuració per cada estratègia estudiada, en termes de precisió. S'hi mostra el valor de $\%AR$ i el nombre mitjà d'operacions per a la recuperació, $\#Op$. També s'hi inclou el coeficient de correlació lineal ρ, entre els valors de $\%AR$ i $\#Op$ pel conjunt de configuracions estudiades. Els problemes de prova apareixen separats en dos grups, en funció de si ρ és > 0.5 o < 0.5.

Dit d'una altra manera: valors elevats de ρ asseguren un pitjor comportament d'aquells algorismes basats en una reducció del número de comparacions en la fase de recuperació de la memòria de casos. En aquests problemes, per tant, no serà possible l'objectiu inicialment proposat per aquestes variacions del CBR: reduir el temps de càlcul d'una manera significativa, tot mantenint la mateixa precisió que s'obté utilitzant un sistema CBR sense clústerització de la memòria de casos.

4.4 Qualitat del SOM i mètriques de complexitat

L'anàlisi fet a l'apartat anterior permet, a partir del coneixement del paràmetre ρ, determinar si les diferents variants del SOMCBR estudiades faran possible el comportament buscat: un manteniment de la precisió respecte el CBR, amb un menor nombre d'operacions a realitzar. No obstant, cal tenir en compte que ρ és un paràmetre calculat a partir dels valors de $\%AR$ i $\#Op$, obtinguts com a resultat de l'aplicació del SOMCBR sobre el problema de prova corresponent.

Per tant, caldrà discutir l'existència de magnituds que es puguin calcular prèviament a l'aplicació dels algorismes, i que aportin una informació equivalent a la que es resumeix amb el factor ρ.

L'objectiu inicial d'allò exposat en aquest apartat és establir quines magnituds permeten determinar en quins casos l'aplicació d'un cert algorisme pot aportar resultats interessants (en aquest cas, respecte la bondat de la classificació realitzada). Si s'arriba a la conclusió a partir de mesures obtingudes un cop aplicat el mètode, l'objectiu no s'assoleix, doncs no té sentit haver d'aplicar l'algorisme per discutir sobre la viabilitat d'aplicar-lo. Per tant, convé determinar de quina manera es pot arribar a conclusions similars a partir de mesures obtingudes *a priori*.

4.4.1 Mesura de la qualitat d'un mapa auto-organitzatiu

Una primera opció, en el cas del SOMCBR, és estudiar directament alguna mesura que il·lustri la qualitat del mapa elaborat a l'etapa del SOM. A mode de context, cal dir que els mapes de Kohonen o mapes auto-organitzatius (*Self Organization Mapping*, SOM) [85] són una de les estratègies no supervisades de més vigència en l'àmbit de les xarxes neuronals, a l'hora de definir agrupacions o clústers. Obtenir un SOM implica projectar les dades d'un espai a un altre tot reduint-ne la dimensionalitat, a partir d'un procés d'entrenament no supervisat.

El que s'aconsegueix d'aquesta manera és facilitar la comprensió de les dades i, si la dimensionalitat final és menor o igual que 3, facilitar-ne també la visualització. A més d'aquestes, el procés compleix tot un conjunt de propietats:

- Preserva la topologia original de les dades
- Proporciona bons resultats per espais originals de dimensions elevades
- No es perd la informació d'un classe amb pocs casos en el problema de prova
- S'ajusta de manera autònoma i no supervisada

Tenint en compte que per qualsevol variant del SOMCBR que s'estudia es fixen les mides del mapa (en els resultats mostrats, 3×3, 4×4 i 5×5), s'intentarà obtenir una mesura de qualitat de la clústerització realitzada, un cop determinada aquesta mida.

Per fer-ho, tinguem en compte els nodes del mapa generat pel SOM més propers per cada instància (*Best-Matching Unit*, BMU), i si els casos representats per aquests nodes pertanyen o no a la mateixa classe que la instància corresponent. Així, la qualitat del mapa generat pel SOM es pot mesurar a partir de l'error de quantització (q_{error}, [86]), definit com l'error promig entre cada instància i el BMU de la mateixa classe. Aquesta mesura té en compte la distribució de les instàncies sobre el mapa generat pel SOM i, per tant, es basa en la topologia d'aquest. Un valor proper a 0 de q_{error} indica que la clústerització ha donat lloc a una elevada separació de classes, i per tant és d'esperar que el SOMCBR realitzi una bona classificació en un temps menor a la fase de recuperació.

Per determinar la possible utilitat d'aquesta mesura de qualitat del mapa generat pel SOM, es procedeix a comparar els seus valors amb els que pren el paràmetre ρ, definit a l'apartat anterior. A la figura 4.3 es representen els valors de q_{error} respecte els valors de ρ pels problemes de prova *Iris, Heart-Statlog, Glass, Breast Cancer Wisconsin, Vehicle, Ionosphere, Sonar, Tao, μCa, MIAS-Birads* i *MIAS-3C*.

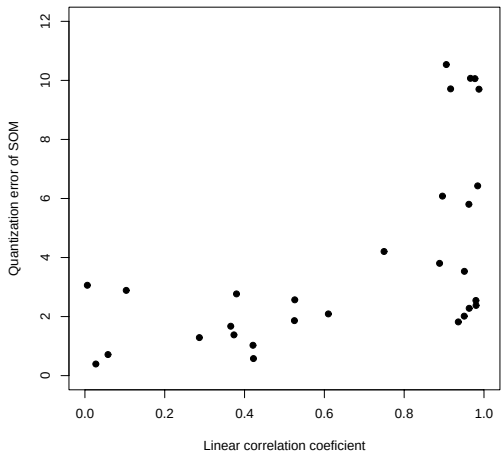

Figura 4.3: La figura mostra els valors de q_{error} (*quantization error of SOM*) respecte els valors de ρ (*linear correlation coeficient*) pels problemes de prova *Iris, Heart-Statlog, Glass, Breast Cancer Wisconsin, Vehicle, Ionosphere, Sonar, Tao, μCa, MIAS-Birads* i *MIAS-3C*, i les diferents configuracions de mapa testejades (3×3, 4×4 i 5×5).

Es pot observar com existeix una certa relació entre ambdós valors: tots els casos en què $\rho \lesssim 0.6$ els valors de q_{error} són baixos (concretament, < 4), i sempre que $q_{error} \gtrsim 4$ implica valors elevats de ρ. Aquests resultats són coherents amb l'esperat: en aquells casos en què la qualitat del mapa gen-

erat no és la millor possible, la clústerització no permet mantenir la precisió reduint el número de comparacions i, per tant, ρ és elevat. És a dir, el resultat del SOMCBR no serà el desitjat, doncs la precisió serà extremadament proporcional amb la mida de la memòria de casos utilitzada en la recuperació.

Aquest resultat, però, té dos punts febles: d'una banda, la implicació no es pot treballar en el sentit invers. És a dir, no es pot assegurar que en tots els casos en què el resultat del SOMCBR no sigui el desitjat ($\rho \gtrsim 0.9$), aquest comportament serà predit pel valor de q_{error}, doncs hi ha casos compatibles amb q_{error} baixos. D'altra banda, el problema principal rau en què el càlcul del paràmetre q_{error} implica haver d'aplicar el SOM per generar el corresponent mapa i, per tant, no es compleix del tot el que es desitjava: obtenir una mesura absolutament *a priori* per determinar el futur comportament del SOMCBR, només depenent del problema de prova corresponent.

4.4.2 La utilitat de les mètriques de complexitat

Una alternativa al plantejat fins ara és estudiar la complexitat inherent de cada problema de prova. Centrant-nos en el cas particular que s'analitza, la bondat dels resultats tindrà molt a veure amb la capacitat del SOM de separar l'espai que ocupen els casos coneguts, de tal manera que els clústers obtinguts agrupin casos amb similars característiques.

És possible que això tingui relació amb la complexitat de cada problema, malgrat desconeguem de quina manera afectarà. Sí que és cert que un problema que sigui poc separable pel SOM implicarà un resultat que serà de menor qualitat, doncs al triar el clúster o clústers més significants per realitzar la recuperació es perdrà informació determinant. En canvi, si el SOM fa una bona separació dels casos en clústers, la tria entre aquests en la fase de recuperació aportarà resultats similars als del CBR, reduint apreciablement el nombre d'operacions en aquesta fase.

A partir d'aquí, la qüestió se centra en com definir la complexitat d'un problema de prova, i en quin és el seu efecte sobre el comportament del SOM. Aquest és un concepte pel qual no hi ha una única definició, ans al contrari: diversos estudis, molt en particular el de T.K.Ho i M.Basu ([78]), han proposat tot un conjunt de magnituds per mesurar aquest concepte, conegudes habitualment com a mètriques de complexitat. Aquest estudi parteix de la base que difícilment existeix una única mesura que pugui caracteritzar globalment totes les causes per les quals un problema pot ser considerat "complex", per la qual cosa és més adient considerar cada problema com un punt en un espai n-dimensional definit per n mètriques de complexitat.

Aquestes mètriques poden classificar-se en tres grups: d'una banda, hi ha aquelles que tracten la separabilitat lineal del problema (L1, L2, L3 en la nomenclatura del citat estudi); aquelles que estan basades en l'estudi dels veïns més propers (N1, N2, N3, N4); i aquelles que treballen sobre la geometria i la topologia del problema (F1, F2, F3, T1, T2). Per simplicitat, es definiran posteriorment de manera més extensa només aquelles que s'utilitzaran per als càlculs realitzats.

Totes elles estan definides per a problemes de classificació de dues classes, essent generalitzables a un problema de més classes de manera no trivial. Per aquest motiu, i també per poder comptar amb un número de problemes de prova major per a l'estudi, habitualment es separa un problema de m classes en m problemes de dos classes, construïts a partir de determinar una de les m classes com a 0 i la resta com a 1. Queden, per tant, problemes de prova del tipus "una classe contra les altres": és classe "A", o no ho és.

4.5 Càlcul de les mètriques de complexitat

L'objectiu d'aquest apartat, i dels que segueixen, és l'estudi de les mètriques de complexitat, des del punt de vista de la seva utilitat per a la determinació *a priori* del bon comportament d'un algorisme sobre un problema de prova. Els càlculs que segueixen estan realitzats sobre la col·lecció de problemes utilitzats a [22]. La majoria són extrets del repositori UCI [3] (com *Iris*, *Wine*, *Thyroid*, *Balance*, *Vehicle*, *Waveform*, *Heart-Statlog*, *Ionosphere*, *Wisconsin*, *wbcd*, *wpbc*, *wdbc* i *Pima*), i d'altres són d'ús habitual per part del Grup de Recerca en Sistemes Intel·ligents (com *TAO* o *bpa*). A la taula 4.5 s'hi exposen les seves característiques principals.

D'acord amb el que s'ha dit anteriorment sobre el càlcul de les mètriques de complexitat per problemes de dues classes, aquells en què els casos poden pertànyer a més de dues classes són separats en diversos problemes de dues classes. Així, *Irisc1* és el problema de prova en què les classes 2 i 3 estan agrupades en una sola nova classe, *Irisc2* agrupa les classes 1 i 3, i *Irisc3* ho fa amb les classes 1 i 2. La nomenclatura és generalitzables per tots els problemes de prova utilitzats.

Per tots aquests problemes, s'han calculat els valors de les següents mètriques de complexitat [78]:

- Relació discriminant de Fisher (F1): per un problema amb diversos atributs, es mesura la relació entre la diferència de mitjanes dels valors

Sigles	Prob. prova	Atributs	Classes	Distribució per classes
IR	Iris	5	3	iris-setosa (50), iris-versicola (50), iris-virginica (50)
HS	Heart-Statlog	14	2	absent (150), present (120)
WS	Wisconsin	10	2	benign (458), malign (241)
VE	Vehicle	19	4	0 (212), 1 (217), 2 (218), 3(199)
IO	Ionosphere	35	2	b (126), g (225)
WI	Wine	14	3	1 (59), 2 (71), 3 (48)
TH	Thyroid	6	3	1 (150), 2 (35), 3 (30)
BA	Balance	5	3	B (49), L (288), R (288)
WA	Waveform	41	3	1 (150), 2 (35), 3 (30)
WB	wbcd	10	2	b (458), m (241)
WP	wpbc	34	2	b (151), m (47)
WD	wdbc	31	2	b (357), m (212)
PI	Pim	9	2	0 (500), 1 (268)
TA	Tao	3	2	black (944), white (944)
BP	bpa	7	2	

Taula 4.5: Descripció dels problemes de prova utilitzats per a l'assaig de les diferents propostes de SOMCBR, provinents de diversos repositoris, i dels quals s'han calculat els valors de les mètriques de complexitat definides al text. De cada un d'ells s'indica l'habitual abreviació, el nom complert, el número d'atributs de cada instància (contant-hi la classe a la qual pertanyen), el número de classes, i la distribució de les instàncies per cada classe.

de cada atribut per les dues classes (A i B), i la suma de les variàncies per ambdues classes. Dels valors obtinguts per cada atribut, F1 és igual al màxim d'aquests: si existeix un atribut altament discriminant, el valor de F1 és molt elevat. Habitualment es calcula com:

$$F1 = Max\left(\frac{(\mu_A - \mu_B)^2}{\sigma_A^2 + \sigma_B^2}\right)$$

- Volum de la regió de solapament(F2): mesura el solapament entre els valors dels atributs per les dues classes (A i B), a partir d'un percentatge de volum solapat respecte l'ocupat. Si es suposa un problema amb N atributs, el càlcul és per tantes dimensions com atributs:

$$F2 = \prod_{i=1}^{N} \frac{Min(max(C_{A,i}), max(C_{B,i})) - Max(min(C_{A,i}), min(C_{B,i}))}{Max(max(C_{A,i}), max(C_{B,i})) - Min(min(C_{A,i}), min(C_{B,i}))}$$

on $max(C_{A,i}$ representa el valor màxim de l'atribut i per als casos de classe A, i així successivament.

- Eficiència dels atributs (F3): descriu la contribució de cada atribut a la separabilitat entre les dues classes possibles. Tenint present que si en algun atribut no hi ha solapament per les dues classes aleshores el problema és perfectament separable, F3 es calcula com el percentatge de casos que cauen fora de la regió de solapament, en el cas de l'atribut que més contribueixi a la separabilitat.

- Punts a la frontera (N1): partint del càlcul d'un arbre mínim de connexions (*minimum spanning tree*, MST, [87]), que connecta cada cas amb el seu veí més proper en l'espai determinat pels atributs del problema, N1 es calcula com el percentatge de casos que estan connectats amb altres de diferent classe, respecte el total de casos existents en el problema.

- Relació mitjana intra/inter (N2): de manera similar a l'anterior, es basa en un MST. En aquest cas, N2 és el quocient entre el valor mig de les distàncies intra-classe (entre els veïns més propers de la mateixa classe) i el valor mig de les distàncies inter-classe (entre els veïns més propers de diferent classe).

- No-linealitat (N4): l'objectiu és la mesura de l'error d'un classificador sobre casos creats artificialment a partir de la combinació lineal dels casos del problema, amb pesos aleatoris. N4 és l'error si el classificador utilitzat és un de veïns més propers (NN).

- Cobertura de l'espai (T1): relacionat amb conceptes topològics ([88]), avalua si els casos tendeixen a agrupar-se en clústers o bé a estar extesos a través de primes estructures, en l'espai definit pels atributs.

- Punts per dimensió (T2): també de caràcter topològic però molt més simple que l'anterior, intenta donar una mesura de la dimensionalitat real del problema, i es calcula com el quocient entre el número de casos i el número d'atributs.

Totes aquestes mesures són inherents al problema de prova i, per tant, independents de l'algorisme assajat. Els resultats obtinguts es mostren a la taula 4.6, junt amb d'altres magnituds que es definiran a continuació.

Problema de prova	Mètriques de complexitat N1	N2	N4	F1	F2	F3	T1	T2	%R	p-value
Waveform c1	0.24	0.86	0.10	1.32	0.89	0.23	1.00	125.00	89.2	0.00
Vehicle c1	0.12	0.42	0.39	1.12	0.60	0.46	0.99	47.00	86.9	0.00
Vehicle c4	0.09	0.54	0.59	0.38	0.72	0.22	1.00	47.00	87.7	0.00
Balance c2	0.20	0.62	0.67	0.38	1.00	0.00	0.90	156.25	89.7	0.00
Waveform c2	0.27	0.90	0.20	1.17	0.94	0.15	1.00	125.00	89.7	0.01
Pim	0.44	0.84	2.49	0.57	0.84	0.01	0.99	96.00	87.9	0.03
Wpbc	0.42	0.91	1.68	0.47	0.79	0.18	1.00	6.00	82.5	0.03
Waveform c3	0.23	0.85	0.07	1.41	0.89	0.24	1.00	125.00	89.3	0.03
Balance c3	0.20	0.62	0.67	0.38	1.00	0.00	0.90	156.25	89.5	0.04
Tao	0.07	0.16	2.16	1.39	0.69	0.36	0.40	944.00	81.8	0.06
Wdbc	0.07	0.56	0.00	3.39	0.61	0.52	0.99	18.97	80.2	0.09
Wbcd	0.06	0.34	0.12	3.46	0.86	0.12	0.70	77.6	86.9	0.00
Vehicle c2	0.37	0.71	1.67	0.18	0.66	0.04	0.99	47.00	81.9	0.11
Vehicle c3	0.37	0.74	2.36	0.17	0.64	0.06	0.99	47.00	82.5	0.11
Bpa	0.58	0.91	2.66	0.05	0.65	0.03	1.00	57.50	52.6	0.17
Heart-Statlog	0.37	0.67	0.31	0.75	0.88	0.01	1.00	20.77	87.1	0.19
Balance c1	0.21	0.65	1.07	0.00	1.00	0.00	0.89	156.25	89.00	0.21
Wisconsin	0.06	0.33	0.48	3.59	0.86	0.12	0.80	77.67	84.5	0.33
Ionosphere	0.23	0.63	0.24	0.61	0.00	0.19	0.95	10.32	64.00	0.41
Iris c2	0.01	0.10	0.00	16.65	0.00	1.00	0.89	37.50	56.30	0.00
Thyroids c2	0.06	0.23	0.00	3.50	0.10	0.81	0.68	43.00	52.80	0.01
Thyroids c1	0.05	0.23	0.00	2.48	0.16	0.85	0.73	43.00	51.4	0.02
Iris c1	0.09	0.17	0.00	3.89	0.28	0.75	0.45	37.50	60.7	0.04
Wine c1	0.05	0.43	0.00	5.39	0.45	0.72	0.99	13.69	68.6	0.05
Wine c2	0.07	0.49	0.00	4.23	0.46	0.76	0.99	13.69	67.9	0.05
Thyroids c3	0.10	0.31	1.55	0.25	0.26	0.19	0.84	43.00	54.2	0.08
Iris c3	0.10	0.21	1.67	0.67	0.43	0.56	0.82	37.50	60.9	0.08
Wine c3	0.12	0.57	0.47	2.33	0.60	0.58	0.99	13.69	65.3	0.09

Taula 4.6: Valors de les mètriques de complexitat calculades, del percentatge promig de reducció del cost computacional a la fase de recuperació ($\%R$) i de l'invers de la probabilitat de rebuig de la hipòtesi nul·la entre la precisió obtinguda per CBR i per les diferents configuracions de SOMCBR (p). La línia horitzontal diferencia els problemes de prova de tipus 1 i 2, tal i com s'ha definit al text.

4.6 La complexitat com a mesura prèvia

Un cop calculades totes aquestes mètriques de complexitat de cada un dels problemes de prova, i per tal d'analitzar els resultats obtinguts, és essencial determinar correctament què significa que el SOMCBR (o qualsevol altra sistema classificador utilitzat) té un bon o un mal comportament. Posteriorment, s'analitzarà si la complexitat és una magnitud relacionada amb aquest comportament.

4.6.1 La bondat de l'algorisme i ρ

Fins ara s'ha utilitzat el valor de la magnitud ρ, calculat després d'aplicar l'algorisme sobre el problema de prova: per valors elevats, significa que un bon resultat en termes de precisió només és possible utilitzant la pràctica totalitat de la memòria de casos, amb el consegüent cost computacional en l'etapa de recuperació. Per tant, són valors alts de ρ els que s'associen amb mals comportaments del SOMCBR.

No obstant això, aquest plantejament deixa fora alguns casos que podrien ser d'interès: aquells en els quals, malgrat reduir la precisió de manera significativa, el guany computacional sigui prou important com per reportar un benefici, factor que no es veu contemplat en el valor de ρ.Depenent de l'objectiu i l'aplicació de l'algorisme, es pot sacrificar la precisió en la classificació, per exemple, a canvi d'una molt menor càrrega computacional.

Dit d'una altra manera, l'estudi del comportament d'un algorisme requereix més informació que l'aportada per ρ: és quelcom més complex, que no depèn només d'un percentatge d'error, sinó també de consideracions sobre el número d'operacions a realitzar. A tal efecte, es defineixen dues noves variables: d'una banda, el percentatge de reducció d'operacions en la fase de recuperació ($\%R$). Valors elevats de $\%R$ impliquen una elevada reducció del cost computacional o, dit d'una altra manera, la necessitat d'utilitzar només una petita part de la memòria de casos per a la recuperació.

D'altra banda, es defineix p com l'invers de la probabilitat de rebutjar la igualtat de comportament entre el CBR i les diferents estratègies del SOMCBR. Aquesta "igualtat de comportament" és el que al capítol 6 es definirà com a hipòtesi nul·la, en el moment d'explicar com afecta al plantejament del problema la manera de presentar les hipòtesis. Valors elevats de p indiquen una baixa probabilitat que els resultats obtinguts per CBR i per SOMCBR siguin significativament iguals. Per tant, la primera variable aporta informació sobre la mida de la memòria de casos utilitzada (relacionat amb $\#Op$), i la segona sobre la bondat del resultat en precisió (relacionat amb $\%AR$).

D'aquesta manera es pot definir *ad hoc* el que es consideri un problema en què el SOMCBR té un mal comportament: aquell en què els valors de $\%R$ i de p siguin baixos. És a dir, quan la reducció en el temps de càlcul no sigui important, i quan la reducció en la precisió sigui estadísticament significativa (i per tant existeixi una elevada probabilitat de rebutjar la igualtat de comportament entre el CBR i el SOMCBR).

Aquests problemes de prova són els que en direm de *tipus 2*, mentre que els de *tipus 1* seran aquells en què es complirà com a mínim una de les dues

condicions: o bé existeix una reducció apreciable del temps de càlcul a la fase de recuperació (valors elevats de $\%R$), o bé la probabilitat d'igualtat de precisió en el resultat del SOMCBR respecte el CBR és alta (valors no baixos de p). O bé, és clar, totes dues coses.

A la taula 4.6, que s'ha introduït anteriorment, es mostren també els resultats pel conjunt dels 28 problemes de prova utilitzats. A banda dels valors per les mètriques de complexitat calculades, apareixen els valors corresponents a aquests dos paràmetres definits: $\%R$ i p. D'acord amb els resultats obtinguts i la significació real de cada un dels problemes, en el cas que es presenta es consideraran problemes de *tipus 2* aquells en què $p < 0.10$ i $\%R < 70$, la qual cosa té lloc simultàniament pels problemes de prova *Iris*, *Wine* i *Thyroid* (per qualsevol de les seves classes), tal i com es pot observar a la citada taula.

4.6.2 Relació de la complexitat amb $\%R$ i p

Els dos paràmetres definits ($\%R$ i p) segueixen essent, malgrat la seva utilitat en la determinació del comportament d'un algorisme, magnituds calculades *a posteriori* de l'aplicació de l'algorisme. Per tant, resta determinar si existeix alguna relació entre els resultats de les mètriques de complexitat (càlcul *a priori*) i aquestes dues variables. Aquestes relacions es troben a partir de la següent anàlisi.

En primer lloc, és interessant estudiar la dependència entre els valors de precisió que el CBR o la millor configuració del SOMCBR assoleixen per a cada problema de prova ($\%AR$), i el valor de la mètrica N1. Tal i com es pot observar a la figura 4.4, valors elevats de N1 impliquen valors menors de precisió; és a dir, en aquells problemes de prova més complexos en el sentit de N1 l'algorisme classificador mostra menor capacitat de classificació. La relació, a més, té un alt component lineal, amb un coeficient de correlació lineal igual a -0.98 (on el signe negatiu indica el pendent negatiu de la recta de regressió entre aquestes dues magnituds).

Aquesta mètrica està altament correlacionada amb el valor que pren N2 (ambdues basades en els veïns més propers), tal i com s'observa a la figura 4.4. És per aquest motiu que, de cara a potenciar l'efecte d'aquestes mètriques a l'hora d'analitzar l'espai de complexitat, es defineix una nova mètrica N_{12} com el producte $N1 \times N2$.

Malgrat els resultats obtinguts, anàlisis com els efectuats a la figura 4.4 encara no aporten una relació clara entre mètriques de complexitat i les

mesures que es proposen per determinar la tipologia del problema de prova ($\%R$ i p). Per fer-ho, cal primer estudiar el comportament d'una respecte l'altre, i posteriorment el de cada una d'elles respecte les mètriques. Això ha de permetre escollir un conjunt de mesures de complexitat que converteixin en separable la regió d'aquest espai on es trobin els problemes de prova de *tipus 1* i els de *tipus 2*.

La relació entre $\%R$ i p es pot observar a la figura 4.5, on les línies perpendiculars que apareixen marquen els valors límits per ambdues magnituds. Aquesta gràfica visualitza les definicions que s'han fet anteriorment sobre el comportament del SOMCBR, per bé que sobre les variables obtingudes *a posteriori* de l'aplicació de l'algorisme: aquells problemes de prova pels quals el SOMCBR no té un bon comportament (*tipus 2*) són aquells pels quals trobem valors baixos de $\%R$ (< 70) i de p (< 0.1).

A la figura 4.6 es mostren les relacions entre aquestes dues magnituds i la mètrica N_{12}, introduïda anteriorment: en la primera de les dues, per exemple, no es manté la separabilitat observada a la gràfica 4.5, cosa que sí passa a la segona. En tot cas, N_{12} apareix com una magnitud d'interès per als nostres propòsits, doncs els problemes de *tipus 2* ocupen només regions amb valors molt propers a 0 d'aquesta magnitud, cosa que també passa p p i $\%R$, degut a la pròpia definició.

Tant N1 com N2 són mètriques basades en les relacions amb els veïns més propers. Per completar l'anàlisi, es proposa utilitzar una mètrica que contingui informació topològica i que, a més, mostri una elevada separabilitat entre els tipus *1* i *2* a l'hora de relacionar-ho amb els factors $\%R$ i p. Aquesta mètrica pot ser $F3$, tal i com es mostra a la figura 4.7, on s'observa un comportaas problemes de *tipus 2* en sentit invers que amb N_{12}: la seva tendència és cap als valors elevats de $F3$.

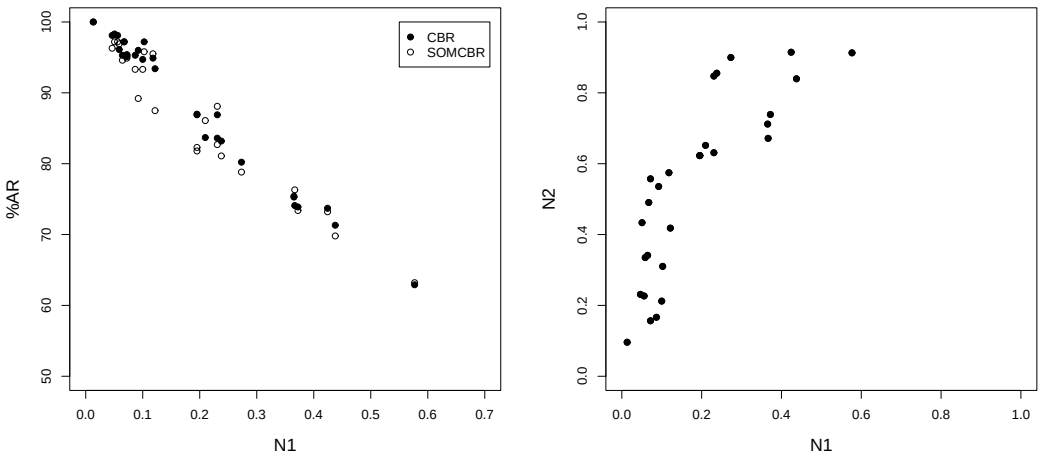

Figura 4.4: En la figura de l'esquerra es mostra la relació altament lineal entre la precisió en la classificació i el valor de la mètrica N1, que mesura el percentatge de punts que hi ha a la frontera de cada classe. Els valors de $\%AR$ són els obtinguts pel CBR i per la millor configuració del SOMCBR, en termes de precisió. A la figura de la dreta, es mostra la relació entre les mètriques N1 i N2, amb un comportament proporcional.

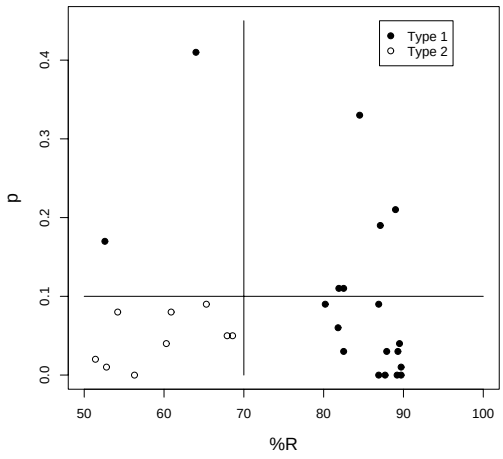

Figura 4.5: Relació entre els valors de p i $\%R$ pel conjunt de problemes de prova estudiats. Les rectes perpendiculars marquen les regions definides pels valors $p = 0.1$ i $\%R = 70\%$.

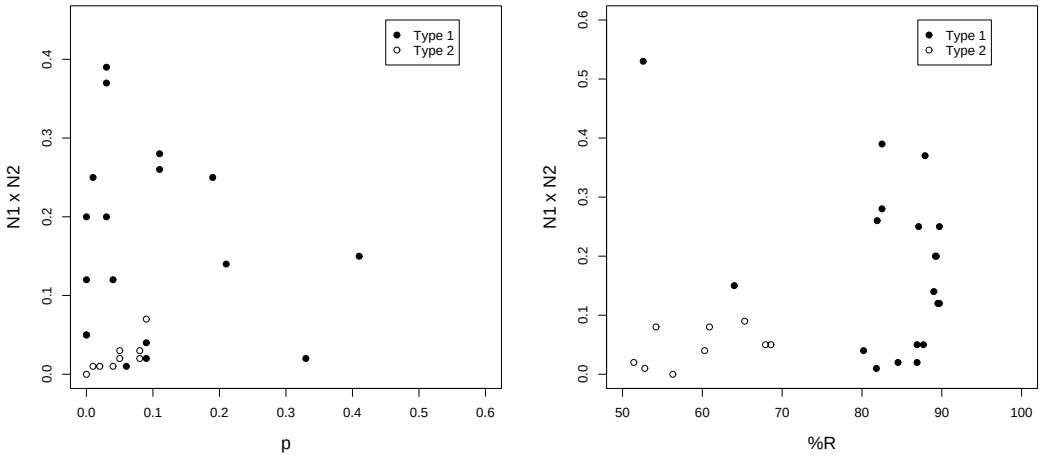

Figura 4.6: En ambudes figures es mostra la relació entre el producte de mètriques $N1 \times N2$ i les variable p i $\%R$. Respecte el primer valor (gràfica de l'esquerra), es perd la separabilitat, que es manté respecte $\%R$ (gràfica de la dreta), tot i observant que els problemes de prova de *tipus 2* es concentren a la regió més propera de l'origen.

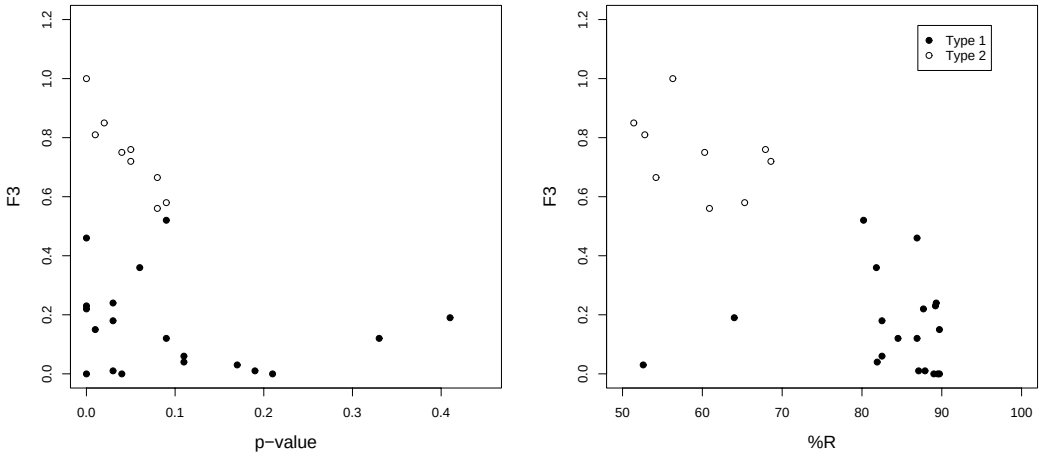

Figura 4.7: En ambudes figures es mostra la relació entre la mètrica **F3** i les variable p i $\%R$, manifestant-se la capacitat de **F3** per a la separació dels tipus diferents de problemes de prova.

4.6.3 Espai de complexitat separable

El conjunt de mètriques estudiades, i les propietats de separabilitat que apareixen en les figures 4.6 i 4.7, suggereixen un conjunt de tendències seguides per la pràctica totalitat dels problemes estudiats:

- Els problemes de prova amb valors alts de $\%R$ apareixen en regions amb valors de F3 baixos.
- Els problemes de prova amb valors elevats de F3 tenen valors de p i de $\%R$ propers als mínims.
- Els valors baixos de $\%R$ estan altament correlats amb amb valors propers a 0 de N_{12}.

Aquests resultats ens aporten un espai de complexitat per separar els problemes de prova d'ambdós tipus, generat per les variables N_{12} i $F3$. A la figura 4.8 es mostra la representació dels problemes estudiats en el citat espai de complexitat. Les conclusions venen marcades per les dues rectes perpendiculars que s'hi dibuixen: els valors de $F3 > 0.55$ i $N_{12} < 0.10$ determinen un rang en el qual els problemes de prova sempre són de *tipus 2* i, per tant, obtindran mals resultats en l'aplicació del SOMCBR, en el sentit que s'ha definit abans.

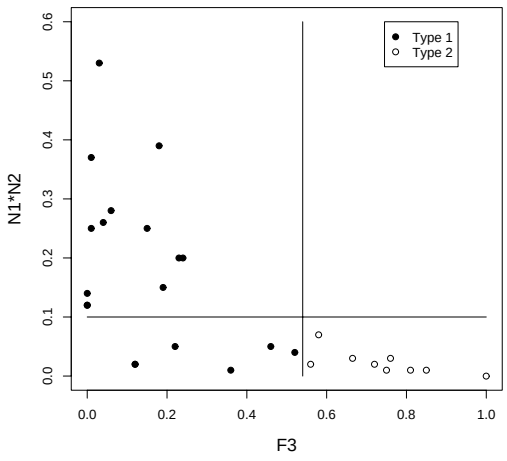

Figura 4.8: Representació dels problemes de prova estudiats en l'espai de complexitat generat per les mètriques N_{12} i $F3$. Les rectes perpendiculars marquen la separació entre la regió on es troben els problemes de prova de *tipus 2* i la resta.

De fet, un valor elevat de F3 implica que els atributs no mostren tots un alt grau de solapament i, per tant, existeix algun o alguns atributs que permeten una alta separabilitat de les classes. En el mateix sentit, un valor proper a 0 de N_{12} implica, com a mínim, un baix nombre de connexions per proximitat entre casos de diferent classe, o bé una alta concentració dels casos d'una mateixa classe. En tot cas, la regió de complexitat ocupada pels problemes de *tipus 2* coincideix amb el que podríem considerar "baixa complexitat", especialment en comparació amb les altres regions de la figura 4.8. Això indica que, en un problema fàcilment separable, la clusterització introduïda pel SOM en els mapes de 3×3, 4×4 i 5×5 sigui excessiva, i no aporti informació de valor en la separació dels casos, ans al contrari.

4.6.4 Conclusió: la complexitat com a *útil* mesura prèvia

Dels apartats anteriors es pot obtenir una conclusió *ad hoc* al problema estudiat, i una altra més generalitzable. La primera és que, pels problemes de menor complexitat en el sentit de N_{12} i $F3$, el SOMCBR no té un bon comportament, d'acord amb el definit per $\%R$ i p. En aquests casos, és recomanable un algorisme classificador més simple o, si es pot assumir el número d'operacions a realitzar, directament el CBR sense cap tractament sobre la memòria de casos.

La segona conclusió és de major abast: la complexitat, mesurada per les mètriques definides i calculada abans de l'aplicació de l'algorisme, aporta informació *a priori* del comportament que tindrà aquest algorisme, sobre un determinat problema de prova. Com a mínim, permet dir si aquest serà de *tipus 1* o de *tipus 2*, seguint les definicions fetes anteriorment.

D'aquí es dedueix el que s'ha exposat al principi com a hipòtesi: l'anàlisi dels resultats obtinguts pels diferents algorismes sobre els problemes de prova pot ser erroni, si no es té en compte les propietats inherents a aquests problemes. En aquest cas, per exemple, el SOMCBR aplicat sobre problemes de *tipus 2* ens duria a unes conclusions totalment distants de les obtingudes amb els problemes de *tipus 1*.

4.7 Determinació de regions de complexitat

El resultat mostrat en l'apartat anterior ens permet considerar que una representació a partir d'aquestes mètriques (N_{12} i $F3$) defineix un cert espai de complexitat que, tenint en compte els aspectes topològics i de frontera que

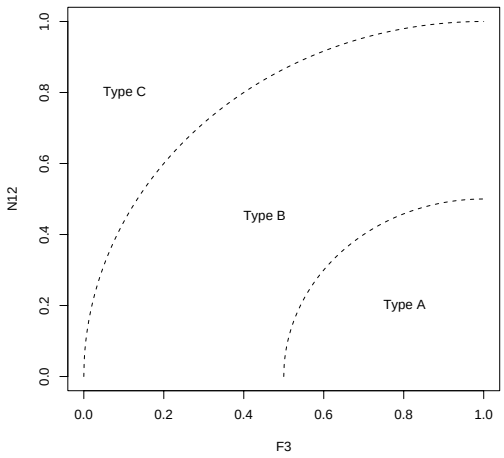

Figura 4.9: Exemple de determinació de les regions de major o menor complexitat, a partir de la distància al punt de menor complexitat del mapa.

cadascuna d'elles inclou, pot ser fàcilment generalitzable.

De fet, en un mapa com l'anterior es poden definir els punts de màxima i mínima complexitat (MCP i mCP), situats a $(0, 1)$ i $(1, 0)$ respectivament, i a partir d'aquests punts caracteritzar unes regions de major o menor complexitat. Aquests punts permeten fer un plantejament més general que l'utilitzat a l'apartat anterior, mostrat a la figura 4.8. Allí, la distribució dels problemes sobre l'espai $(F3, N_{12})$ portava a establir la separabilitat de les regions per valors simples d'ambdues magnituds ($N_{12} < 0.5$ i $F3 > 0.5$).

En canvi, en general es proposa fer un plantejament basat en la proximitat als punts MCP i mCP, amb un esperit similar al de l'optimalitat de Pareto ([89]): la complexitat d'un problema de prova serà menor en tant que més proper estigui el punt que representa al mCP, en un mapa de complexitat donat. Si s'utilitza directament una mesura de distància euclídea, s'obtenen regions com les que es mostren a la figura 4.9, on els límits de la distància venen donats pels valors 0.5 i 1.

Seguint aquest sistema, a la figura 4.10 s'hi representen els problemes utilitzats en l'apartat anterior i detallats a la taula 4.5, d'acord amb el resultat de les mètriques que es mostren a la taula 4.6.

Això permet, a l'hora d'estudiar un determinat problema de prova, situarlo en una determinada regió de complexitat, i mirar de treure conclusions *a priori* sobre l'aplicabilitat d'una determinada variant d'un mètode classifi-

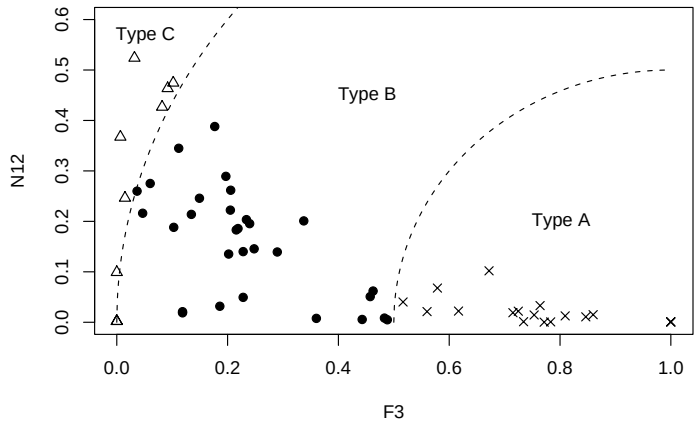

Figura 4.10: Diversos problemes de prova representats en el mapa de complexitat, segons la seva pertinença a una regió de major o menor complexitat.

cador o similar. Per exemple, en els exemples tractats a [2], on s'estudien diverses variants de la clusterització sobre un procés de CBR, aquest estudi aporta conclusions en funció de la complexitat del problema de prova. Així, en aquells casos de poca complexitat la clusterització no és gens efectiva, i existeix una profunda relació de proporcionalitat entre el número d'elements de la memòria de casos utilitzats per la recuperació ($\#Op$) i la precisió obtinguda ($\%AR$). El coeficient de correlació lineal entre ambdues magnituds és en els problemes de la *regió A* del 0.96. Aquesta proporcionalitat es rebaixa pels casos de complexitat mitja (*regió B*), i esdevé molt menor pels problemes amb una major complexitat. En aquests darrers casos es trenca definitivament la proporcionalitat.

Aquest exemple serà tractat amb detall a l'apartat 5.2, quan es discuteix com poder representar de manera simple el comportament d'un número M elevat d'algorismes, assajats sobre una col·lecció d'N problemes de prova, essent N igualment elevat.

4.8 Resum

En aquest capítol s'han presentat diversos casos per mostrar la veracitat de la hipòtesi inicial: l'estudi *a priori* de les característiques dels N problemes de prova, sobre els quals s'assagen els M algorismes, és condició necessària

per a evitar errors en la interpolació dels resultats que s'obtenen.

Per fer-ho, s'ha començat estudiant diverses variants del raonament basat en casos (CBR), basades en la clusterització de la memòria de casos. Concretament, s'ha analitzat la relació entre la precisió obtinguda a l'aplicar els algorismes ($\%AR$) i el número d'operacions a realitzar a la fase de recuperació ($\#Op$), per cada un dels problemes de prova tot definint la magnitud ρ que en determina la linealitat de la relació. Aquesta variable ha permès diferenciar entre els problemes de prova en què no hi ha cap relació de proporcionalitat entre $\%AR$ i $\#Op$ (aquells amb ρ proper a 0, on és possible mantenir la precisió obtinguda pel CBR a partir d'una estratègia que redueixi el nombre d'operacions a realitzar en la fase de recuperació), d'aquelles en què la relació entre ambdós variables és fortament lineal (ρ proper a 1, i per tant no és possible mantenir la precisió reduint el nombre d'operacions a realitzar).

No obstant això, ρ té el gran inconvenient que, per calcular-se, cal haver aplicat totes les variants de l'algorisme sobre el problema de prova pel qual es calcula. En tot cas, els diferents valors obtinguts han posat de manifest que no tots els problemes es comporten igual davant un mateix algorisme.

A continuació, s'ha determinat què es considera, en el cas estudiat, un mal comportament de l'algorisme. La definició de dues variables ($\%R$ i p) ha permès definir els problemes de prova de *tipus 2*: aquells en què els valors de $\%R$ i de p siguin baixos. És a dir, aquells en què la reducció en el temps de càlcul pel SOMCBR no és important i, a més, la reducció en la precisió és estadísticament significativa (i per tant, existeix una elevada probabilitat de rebutjar la igualtat de comportament entre el CBR i el SOMCBR). Ambdós propietats són identificables, en aquests exemples, com a indicadors d'un mal comportament de l'algorisme.

Finalment, s'ha procedit a la qüestió clau: existeixen algunes magnituds que es puguin calcular *a priori*, sense haver d'aplicar els algorismes, que permetin determinar els problemes de *tipus 2*? Una determinada combinació de mètriques de complexitat ha aportat la resposta, permetent a més una doble conclusió.

D'una banda, s'ha posat de manifest la possibilitat de construir un mapa de complexitat en el qual apareixen separats aquells problemes de *tipus 2* de la resta: és a dir, abans de l'aplicació de l'algorisme es pot predir quin en serà el comportament, d'acord amb la definició de $\%R$ i p.

D'altra banda, aquesta diferència de comportament dels problemes respon a la qüestió plantejada a l'inici: no és possible un estudi de la bondat d'uns algorismes sense estudiar les propietats inherents dels problemes de prova

sobre els quals s'assajaran. Aquesta darrera conclusió és generalitzable, més enllà dels exemples presentats: seran diferents les definicions de "bon comportament" pels algorismes, i segurament també el mapa de complexitat que aporti separabilitat de comportament, però queda clar que no serà possible l'estudi de la bondat d'un algorisme sense abans realitzar aquest estudi previ dels problemes de prova sobre els quals s'assaja.

Part III

Metodologies de comparació

Capítol 5

Comparació i representació dels resultats

"**Test** *ESTAD* Donada una mostra de grandària n,
$(x_1, ..., x_n)$, formada per n realitzacions
o observacions independents d'un cert fenomen o experiment,
mètode que permet de decidir si una hipòtesi estadística que hom ha fet
sobre el model probabilístic del fenomen és correcta o no ho és"
Gran Diccionari de la Llengua Catalana

En el capítol anterior, s'ha demostrat la necessitat d'estudiar les propietats dels problemes de prova amb els quals s'assagen els algorismes a estudiar, doncs aquest estudi esdevé un element determinant per evitar conclusions errònies. Aquestes conclusions sorgiran de l'anàlisi dels resultats obtinguts sobre la col·lecció de problemes de prova, amb les metodologies que s'exposaran en els propers capítols. Abans d'això, en aquest capítol s'estudiaran els elements previs a aquesta anàlisi: com establir la hipòtesi a comprovar per fer la comparació entre els algorismes, i de quina manera representar els resultats per a la seva interpretació més ràpida i visual.

Així, en l'apartat 5.1 es desenvoluparà la terminologia per a l'aplicació dels test d'inferència estadística, incloent els conceptes d'hipòtesi i significança estadística. El valor d'aquesta darrera magnitud es discutirà d'acord amb els diferents tipus d'error que es poden cometre ("Tipus I" i "Tipus II").

Un cop la hipòtesi estigui correctament plantejada, en l'apartat 5.2 s'exposaran diferents propostes per a la representació gràfica dels resultats obtinguts pels algorismes sobre les problemes de prova, resultats que permeten

acceptar o rebutjar les hipòtesis plantejades. L'objectiu és limitar l'ús de les llargues taules de resultats habitualment utilitzades, i obtenir unes primeres conclusions qualitatives, en la mesura del possible, abans de realitzar els test d'inferència estadística.

5.1 Plantejament

En general, com ja s'ha explicat anteriorment, es disposen d'M algorismes que es poden aplicar sobre un conjunt d'N problemes de prova, per tal d'avaluar la seva bondat en realitzar una determinada tasca (classificació, predicció, etc.). Així, de l'algorisme i sobre el problema de prova j es disposa d'una mesura de bondat $X_{i,j}$, calculada per alguns dels procediments estimadors exposats al capítol 3. Aquest plantejament no varia en funció de quina és aquesta mesura de bondat: precisió, error, AUC,...

Idealment, la variable que mesura la bondat de l'algorisme i segueix una distribució de valor mig μ_i, que es calcularia sobre els elements de la població de problemes de prova, que seran un total de N_T:

$$\mu_i = \frac{\sum_{j=1}^{N_T} X_{i,j}}{N_T} \tag{5.1}$$

No obstant això, per raons evidents s'està obligat a treballar amb l'estimador no esbiaixat X_i, definit com

$$X_i = \frac{\sum_{j=1}^{N} X_{i,j}}{N} \tag{5.2}$$

on N és el nombre de problemes de prova de què es disposa. Com en general $N << N_T$, és necessari desenvolupar metodologies d'inferència estadística per conèixer la bondat de l'estimador X_i.

En els problemes en què habitualment es treballa, l'objectiu és comparar els valors de X_i obtinguts per a cada algorisme sobre els N problemes de prova de què es disposa, i així determinar allò que interessa: si un determinat algorisme millora el seu comportament respecte els altres, si no hi ha una pèrdua de bondat malgrat els canvis introduïts, etc. Degut a què es treballa sobre estimadors de les magnituds μ_i, els resultats porten implícits la necessitat de preveure un marge de confiança o significança, com es veurà continuació.

5.1.1 Hipòtesis del problema

En inferència estadística es treballa habitualment sobre els test d'hipòtesis, que poden ser acceptades o rebutjades. Una hipòtesi estadística és una predicció o enunciat sobre la relació entre dues o més distribucions, per exemple. El seu test es realitza a partir dels resultats obtinguts sobre els N elements de la població, que en el nostre cas són els problemes de prova.

En el desenvolupament dels test estadístics que s'exposaran, es treballarà sempre amb dues hipòtesis: la hipòtesi nul·la, representada per H_0, i la hipòtesi alternativa, representada per H_1. La hipòtesi nul·la és un enunciat de no diferència entre les magnituds comparades, i la hipòtesi alternativa representa un enunciat que afirma l'existència de diferència entre les magnituds comparades. Tenint en compte el tipus de problemes que s'estudien, habitualment es voldrà demostrar la diferència de bondat entre dos algorismes o estratègies dissenyades, amb la qual cosa normalment l'interessant serà demostrar que es pot rebutjar H_0, o bé que es pot acceptar H_1 ([90]).

Seguint amb la notació habitual, i suposant que s'estudia un test per a la comparació del comportament de dos algorismes A_1 i A_2 sobre una col·lecció de problemes de prova, la bondat dels quals ve representada per les magnituds X_1 i X_2, respectivament, es plantejaran les hipòtesi nul·la i alternativa com:

$$H_0 : X_1 = X_2$$

$$H_1 : X_1 \neq X_2$$

En principi, la hipòtesi alternativa serà no-direccional (també referida sovint com de "doble-cua"). És a dir, la hipòtesi alternativa no té implicació sobre el sentit de la diferència. No obstant això, en alguns problemes pot ser interessant plantejar la hipòtesi alternativa com a direccional (d'una cua), tenint en compte els canvis que implicarà en el càlcul del corresponent test estadístic. Les dues opcions són:

$$H_1 : X_1 > X_2$$

$$H_1 : X_1 < X_2$$

La utilització d'hipòtesis alternatives direccionals està relacionada, en principi, amb el coneixement que a priori es té del problema, o bé amb allò que es vol demostrar. De fet, no hi ha un consens ampli sobre quan utilitzar un tipus o altre d'hipòtesis, i en funció de l'obra consultada es poden trobar arguments en favor d'una i altra opció (direccional o no-direccional, veure

[19]). Donat l'efecte de l'elecció sobre el test estadístic corresponent, l'important és determinar-ho a priori i mantenir el mateix criteri en tot l'anàlisi. Habitualment, s'utilitza l'opció no-direccional i els mateixos valors de X_i determinen, si és el cas, la direcció de la hipòtesi vàlida. Aquesta serà la via d'anàlisi utilitzada en tot aquest treball, doncs l'objectiu habitual és determinar si existeix o no una diferència significativa entre els algorismes comparats, més enllà del sentit d'aquesta diferència, que esdevé prou evident pels mateixos valors de la variable estudiada.

5.1.2 Significança estadística i errors

L'aplicació del mètode d'inferència estadística corresponent per al test de la hipòtesi nul·la (H_0) dóna com a resultat un valor per a l'estadístic utilitzat. Aquest valor s'interpreta a partir de la comparació amb valors crítics: aquells valors llindar, establerts segons les condicions del problema, que permeten rebutjar o no H_0. Els valors crítics han estat calculats prèviament i es troben en taules publicades a la majoria de textos estadístics referenciats en aquest treball.

L'estadístic s'obté a partir dels resultats de l'aplicació dels algorismes sobre el conjunt de problemes de prova. Cal tenir en compte que el conjunt utilitzat per avaluar els algorismes no conté infinits elements, ni tan sols tots els elements possibles de l'univers de problemes de prova (malgrat aquests hagin estat triats a l'atzar d'entre aquesta població). Per tant, com en tot càlcul estadístic fet sobre una mostra, el resultat obtingut pot ser fruit en part de l'atzar, i dur a conclusions errònies: ja s'ha deixat clar al capítol 3 que els valors amb què es treballa són estrictament estimacions de les magnituds que determinen la bondat dels algorismes. La significança estadística marca fins a quin punt s'està disposat a tolerar aquest possible error.

Per exemple, suposem que es detecta una diferència entre els algorismes A_1 i A_2, a partir dels valors X_1 i X_2, i determinada pel valor de l'estadístic z. Es considerarà que aquesta diferència és estadísticament significativa a nivell α si hi ha una probabilitat menor que α de trobar aquesta diferència per atzar, quan realment no existeixi. D'acord amb això, com menor sigui el valor d'α més gran caldrà que sigui la diferència detectada per poder rebutjar la hipòtesi H_0, i major el valor de z obtingut. Les taules de valors crítics contenen, per cada nivell α de significança, el valor mínim de z que cal obtenir per afirmar que la diferència entre X_1 i X_2 és significativa, i per tant rebutjar H_0.

De vegades, també s'interpreta aquest resultat a l'inrevés: a partir de l'es-

tadístic z obtingut, s'intenta determinar el nivell de significança p a partir del qual es podria rebutjar H_0. Si aquest p és menor que el llindar de confiança α que es considera en el problema, es rebutja H_0. En cas contrari, es diu que no hi ha suficients arguments per rebutjar H_0, i per tant cal acceptar l'enunciat proposat per aquesta hipòtesi nul·la.

Habitualment s'accepta el nivell $\alpha = 0.05$ com un bon llindar de significança per als treballs científics, tot i que de vegades es poden obtenir conclusions encara més fortes amb $\alpha = 0.01$. A banda, alguns autors alerten de vegades sobre la diferències entre la significança estadística i la *significança pràctica*.

Aquest matís té sentit, bàsicament, en casos en què els problemes de prova són molts, i això fa que alguns test estadístics portin a resultats significativament diferents a nivell estadístic, si bé la diferència entre les mesures, en valor real o absolut, és extremadament petita, i a la pràctica ho podem considerar com el mateix valor. En els casos en què es treballa en aquest text, és pràcticament impossible trobar-se en aquesta situació: els problemes de prova de què es disposa sovint són escassos, i a més el cost computacional d'obtenir un resultat no és menor. No es tracta, per tant, d'un problema de gran transcendència en aquest treball.

Relacionat amb aquest concepte, i amb el valor d'α que s'utilitza, també es fan servir les notacions d'error "Tipus I" i "Tipus II". L'error "Tipus I" és aquell que es comet quan es rebutja la hipòtesi H_0 essent aquesta vàlida. De fet, la probabilitat de cometre aquest error ve donada pel nivell α de significança amb què es treballa. D'altra banda, l'error "Tipus II" és aquell que es comet quan s'accepta la hipòtesi H_0 essent aquesta falsa: de vegades, es parla també de la potència o capacitat per rebutjar H_0 quan realment existeix una diferència entre A_1 i A_2.[1]

Es pot demostrar que la probabilitat de cometre un error "Tipus II" evoluciona inversament a α i, per tant, si rebaixem el nivell de significança a valors molt propers a 0, per assegurar que no rebutgem hipòtesis H_0 vàlides, cal acceptar el risc d'augmentar la probabilitat d'acceptar una hipòtesi H_0 que no és vàlida. S'ha de tenir present, per tant, que sempre que es faci un esforç per rebaixar l'error "Tipus I" (reduint el valor d'α), de retruc s'està augmentant l'error "Tipus II". Degut a aquesta relació, es considera sovint que $\alpha = 0.05$ és un bon nivell de significança estadística i, de fet, és el nivell habitualment utilitzat en els problemes que s'estudien en aquest àmbit de coneixement.

[1] Aquest concepte s'estudiarà amb més profunditat a l'apartat 8.1, en què s'estudiarà de quina manera avaluar la capacitat d'un test per a determinar una diferència significativa, en cas que aquesta existeixi.

5.1.3 Terminologia per a la classificació dels test

Un cop establerta la condició nul·la H_0 que es sotmetrà a test, a partir dels resultats obtinguts de l'assaig d'M algorismes sobre N problemes de prova, i el valor de la significança α que s'utilitzarà, cal destriar quina metodologia d'inferència estadística es farà servir per a concloure sobre H_0. Les possibilitats són múltiples, depenent de diversos factors: el primer pas, per tant, serà establir una terminologia de les característiques que permeten classificar els test estadístics, per poder presentar a partir del capítol següent aquells que seran d'aplicació a cada un dels problemes que apareixeran.

En primer lloc, cal tenir en compte que en aquest text s'estudiaran aquelles tècniques estadístics d'especial utilitat per als problemes presentats. D'acord amb això, per exemple, es farà referència estrictament a aquells test que treballen hipòtesi sobre la diferència entre mitjanes de les corresponents poblacions, i no pas sobre les seves variàncies. En tot cas, l'estudi de les variàncies i covariàncies permetrà discutir sobre el domini d'ús d'un test, però no sobre la hipòtesi H_0 en si mateixa (com es veurà, per exemple, a l'apartat 6.2.3 en l'estudi del domini d'ús del t-test).

Igualment, cal tenir en compte que habitualment els algorismes s'apliquen sobre un determinat conjunt de problemes de prova que són compartits per tots ells. Per tant, i donat que els resultats s'obtenen sobre els mateixos problemes de prova, els valors obtinguts com a mesura de bondat per a un determinat algorisme (ja sigui l'error de classificació, la precisió, AUC, etc.) no és un conjunt de dades estadísticament independent del resultat obtingut per a un altre algorisme. És a dir, en general s'estudiaran els test que s'apliquen sobre conjunts de dades no independents. Es donarà per suposat, en tots els casos, que la mateixa aplicació de l'algorisme sobre els problemes de prova no afecta al propi algorisme ni tampoc al problema de prova: si fos el cas, pel motiu que sigui, caldria modificar àmpliament les metodologies que s'exposaran als capítols 6 i 7.

A partir d'aquí, hi ha diversos criteris per classificar els problemes que es tractaran. El primer d'ells, que esdevé fonamental i serà el que marcarà la diferència entre els dos capítols que venen a continuació, ve determinat pel número d'algorismes que es volen comparar en la seva aplicació sobre els problemes de prova: parlarem de comparació simple si del que es tracta és de comparar el comportament de dos algorismes entre ells, i de comparació múltiple si la comparació afecta a més de dos algorismes ([91]). Les conclusions que es poden extreure i les tècniques que s'utilitzaran són molt diferents, i també ho és la complicació associada al càlcul del corresponent

estadístic. Això farà que, com veurem, habitualment s'utilitzin només comparacions simples o s'intentin extrapolar aquestes tècniques a problemes de comparació múltiple: estratègia que es demostrarà errònia si el que es vol és discutir realment sobre la bondat d'un conjunt d'M algorismes.[2]

El següent criteri ve determinat per les suposicions que es fan sobre la distribució de la mostra de problemes, íntimament relacionat amb el tipus de dades que s'utilitzen: això determinarà la utilització de test paramètrics o no-paramètrics. Estrictament parlant, un test estadístic paramètric és aquell en el qual es fan un conjunt de suposicions sobre la distribució de la mostra que s'utilitza per avaluar una hipòtesi. En el cas d'un no-paramètric no existeix cap suposició prèvia, tot i que això habitualment això és una mica relatiu: sempre s'assumeix alguna característica sobre el problema de prova, per exemple, per poder construir la tècnica d'inferència.

Sí que serà cert, però, que els test paramètrics assumeixen, sobre els resultats obtinguts, moltes més restriccions que no pas els no-paramètrics. D'acord amb aquest criteri, es parla de test paramètric en casos com el t-test o l'anàlisi de variàncies, en què s'assumeix una distribució normal per la població d'on s'extreu la mostra del problema (entre d'altres condicions), i de test no-paramètrics quan no es suposen aquestes condicions, com els test de Wilkoxon o Frieman ([92]). Tots ells seran estudiats en els propers capítols.

És molt important ser conscients que aquestes suposicions incideixen fortament en la capacitat d'aquests test per a discutir sobre el rebuig o l'acceptació d'H_0: un test no-paramètric, en aplicar-se quan no es poden assumir tot un conjunt de restriccions sobre els resultats obtinguts, redueix la informació de què disposa per al càlcul de l'estadístic que portarà a la discussió d'H_0. En molts casos, la informació numèrica es redueix a qualitativa, establint simples mesures d'ordre relatiu entre X_i per cada j, per exemple.

Aquesta reducció en la informació que s'utilitza implica directament una menor capacitat per al test de la hipòtesi del problema i, per tant, es pot afirmar que en igualtat de condicions un test no-paramètric aporta menys prestacions per a la discussió d'H_0 que un test paramètric[3]. Ara bé, si no es compleixen les condicions marcades per a l'aplicació d'un test paramètric, no existeix alternativa: utilitzar-lo duria a conclusions errònies. Aquestes consideracions es veuran reflectides als protocols d'aplicació que es desen-

[2] Veure, per exemple, l'exposició sobre la matriu de guanys que es fa a l'apartat 6.3.4, i la discussió sobre els errors a què indueix que es fa en l'apartat ??.

[3] Aquestes diferències s'estudiaran pròpiament al capítol 8, a través de conceptes com la potència o la replicabilitat. Les conclusions que s'obtindran confirmaran les consideracions que es fan en aquest capítol.

voluparan al final dels capítols 6 i 7.

5.2 Representació gràfica dels resultats

En l'apartat precedent s'han posat les cases per a l'anàlisi del resultat obtingut: la hipòtesi a discutir, el nivell de significança, la tipologia de test adequat per a l'estudi, etc. Per a poder procedir correctament a realitzar la discussió de la hipòtesi nul·la plantejada, sovint falta encara un altre element: el coneixement qualitatiu dels resultats obtinguts. No serà la mateixa metodologia la que s'aplicarà a la comparació d'M algorismes entre ells, o bé a la comparació d'$M-1$ algorismes introduïts respecte un algorisme ja conegut, per exemple.

Per això, podem afirmar que el coneixement d'algunes magnituds que resumeixin els resultats obtinguts permet plantejar més concretament el problema des d'un bon principi. En aquest apartat s'exposaran diferents formes de representar gràficament el resultat, i es mostrarà amb un cas real els avantatges d'aquests esquemes per sobre de les habituals taules de valors.

5.2.1 El cas univariant

La situació més habitual és aquella en què la bondat de l'algorisme pot ser determinada per un únic valor: la precissió o el percentatge d'error comès, per exemple. En aquest cas, els resultats obtinguts es mostren sempre en una taula de $M \times N$ elements X_{ij}, on cada un d'aquests elements expressen el resultat obtingut per l'assaig de l'algorisme i sobre el problema de prova j. El càlcul del valor mig per a la col·lecció dels N problemes de prova:

$$X_i = \frac{1}{N} \sum_{j=1}^{N} X_{ij} \tag{5.3}$$

permet tenir una primera informació sobre la bondat *mitjana* de l'algorisme A_i, tot i que aquest resultat està massa exposat a efectes com la presència d'*outliers* (com s'analitzarà a l'apartat 6.2.3, en la discusió del domini d'ús del t-test), especialment si es calcula directament sobre els resultats numèrics obtinguts (cas en què es pugui aplicar un test paramètric, i no calgui una reducció de la informació de què es disposa).

Aquesta mesura, però, permet una primera ordenació dels algorismes, i és la que portarà a figures com la 5.1, on s'hi afegeix la informació d'altres magntiuds que es definiran més endavant, com la distància crítica (CD),

definida com la distància mínima d'una variable a partir de la qual es pot assegurar l'existència d'una diferència significativa[4]. Aquest esquema permet determinar quin tipus de comparació es farà i, un cop enllestit el càlcul de l'estadístic i de la distància crítica, representar gràficament el resultat obtingut.

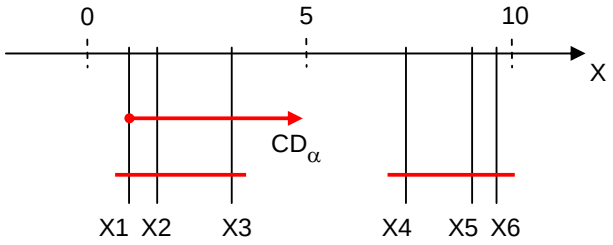

Figura 5.1: Exemple de representació per una magnitud de bondat, X. Els valors $X_1, ..., X_6$ representen sobre l'eix X la bondat dels algorismes $A_1, ..., A_6$, mentre que la distància crítica CD es mesura a partir del resultat de l'algorisme de major bondat, en aquest cas A_1. Aquells en què la distància respecte X_1 és menor que el valor de CD es consideraran significativament equivalents: no es podrà rebutjar la hipòtesi nul·la. Per això apareixen agrupats per un segment horitzontal, al igual que la resta de resultats, que sí tenen un valor significativament diferents al de X_1.

La limitació d'un primer esquema com aquest és evident, però fa possible una primera anàlisi sobre quins algorismes comparar, si n'existeix cap que es pugui utilitzar de control (que és com s'anomena aquell respecte el qual es compara la resta), etc. En els capítols següents s'utilitzarà tot sovint a aquest efecte.

5.2.2 El cas multivariant

En altres ocasions, les magnituds que permeten discutir sobre la bondat d'un algorisme són múltiples. En aquets cas, un problema d'aquest tipus porta a l'anàlisi d'una taula de $M \times N$ elements d'interès (com la precisió o l'error, però també el cost cumputacional o el nombre d'operacions realitzades en una certa etapa de l'algorisme). Si, a més, es vol analitzar el comportament en funció de magnituds *a priori* dels problemes de prova (com es fa a [2] o a

[4]Per un estudi més detallat, veure l'apartat 7.3.4.

[22]), la dificultat per una fàcil visualització dels resultats és encara major, doncs s'estarà interessat en l'estudi dels resultats de manera separada per diversos grups de problemes de prova (en funció de la regió de complexitat que ocupin, per exemple).

En aquesta situació, apareix una important dificultat a l'hora de decidir respecte quin algorisme es comparen la resta, bàsicament perquè cal definir què vol dir "millor": si el número d'indicadors de bondat és 2 o major, aquesta consideració no és evident. Una adequada representació dels resultats facilita aquesta discusió. La proposta que es fa en aquest treball és desevolupar una representació gràfica que permeti, per valors elevats de M i N:

- representar la bondat donada pels valors de X_{ij} sobre els N problemes de prova;
- representar també la dispersió dels seus valors al voltant del valor utilitzat per representar la bondat;
- fer el mateix per una segona magnitud de bondat a analitzar en el problema, representant-hi també la seva dispersió;
- construir aquestes representacions per a les diferents regions de complexitat dels problemes de prova, tal i com s'ha definit a l'apartat 4.7.

A més, sobre aquestes representacions gràfiques es podrà determinar visualment l'equivalència estadística entre els conjunt d'algorismes testejats i, si n'hi ha un que fa el paper de control, determinar aquells que li són estadísticament equivalents, donat un cert valor de significança estadística. Aquest darrer punt ja apareix en els esquemes proposat a l'article de Demsar el 2006 ([9]), però són esquemes massa simples per als casos en què hi ha una major complexitat d'anàlisi: diverses mesures de bondat a tenir en compte, un valor elevat de M, problemes de prova en diferents regions de complexitat, etc.

L'esquema proposat és com el que es veu a la figura 5.2, que a banda d'intentar evitar l'ús de grans taules de resultats, permet una visualització ràpida dels resultats segons dues magnituds relatives a la bondat (X i Y): cada el·lipse representa els resultats obtinguts per l'aplicació d'un algorisme, i té l'origen en el valor utilitzat per representar la bondat dels valors de X_i i Y_i sobre els N problemes de prova. A més, els semieixos de les el·lipses són proporcionals als valors de la dispersió de X i Y per cada algorisme, dades que permeten comparar des d'un altre punt de vista resultats especialment propers en valors mitjos d'aquestes magnituds.

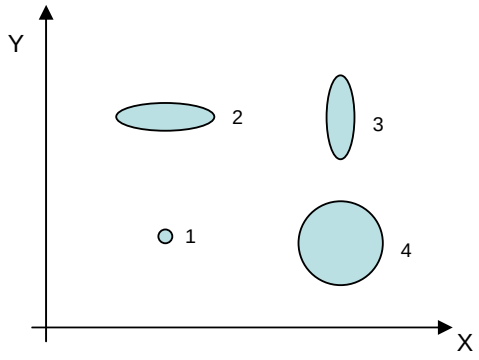

Figura 5.2: Exemple de representació per dues magnituds de bondat, X i Y, producte de l'aplicació de 4 algorismes sobre un conjunt de problemes de prova amb comportaments resultants ben diferents. Tal i com s'explica al text, cada el·lipse representa el resultat d'un algorisme.

Si suposem que la bondat augmenta per valors creixents d'ambdues magnituds, en l'exemple proposat veiem com l'algorisme 3 és millor que el 2, doncs té un bon comportament des del punt de vista de X i també respecte Y (per bé que amb resultats amb una major dispersió al voltant del seu valor mig). L'algorisme 1 és el pitjor de tots, i ho és pràcticament sobre tots els problemes de prova, doncs la dispersió baixa en ambdós eixos indica una molt baixa variabilitat respecte el valor mig. En canvi, l'algorisme 4 és comparable en quant a bondat amb el 2 (mantenen comportaments inversos en cada una de les magnituds), tot i que l'elevat valor de la dispersió en les dues variables el faci poc aconsellable en alguns casos, per la diferència de comportament d'un problema de prova a un altre.

Sigui com sigui, l'esquema és molt simple i permet una ràpida valoració dels algorismes un cop obtingut els resultats sobre la col·lecció de N problemes de prova. A banda, es pot també intentar determinar una mesura única de bondat a partir d'una distància definida des de l'origen (en aquest esquema, el punt de pitjor comportament), d'igual manera com es fa en els mapes de complexitat definits a l'apartat 4.7 (en aquell cas, a partir del punt de mínima complexitat). En funció de les necessitats i les magnituds a analitzar, es poden també invertir el sentit dels eixos o utilitzar-los logarítmics, de cara a reforçar gràficament les diferències. Un bon exemple d'això és el desenvolupat a [2], on la precisió i el número d'operacions en la fase de recuperació són les dues magnituds a comparar entre els diferents algorismes definits.

Pel que fa a la precisió en els resultats, l'ús d'una mesura comparativa entre algorismes té una virtut important, i és que permet determinar sobre la gràfica mateixa l'existència o no de diferències significatives entre els algorismes comparats, com a mínim respecte una de les variables representades. Això és possible tant si s'utilitza una metodologia de comparació paramètrica o com una de no-paramètrica, doncs en ambdós casos es pot calcular una distància crítica tal i com s'ha definit abans (CD, veure més endavant els apartats 7.3.4 i 7.4.2). En el següent exemple es desenvolupa un cas en què s'utilitza aquesta magnitud i l'esquema introduït en aquest apartat.

5.2.3 Un cas pràctic

Un bon exemple per mostrat la bondat de l'esquema proposat és a partir de les dades publicades a [2]. En aquest cas s'utilitza una metodologia de test no-paramètrica, i per tant una mitjana de rangs com a magnitud per a avaluar la bondat en precisió d'un algorisme respecte els altres $M-1$. En aquesta línia, definim R_{ij} com el rang de l'algorisme i aplicat sobre el problema de prova j, respecte els altres $M-1$ algorismes aplicats sobre el mateix problema, de tal manera que $R_{ij}=1$ significa que l'algorisme i és el que dóna millor resultat de tots els M algorismes sobre el problema de prova j ($R_{ij}=M$ implica just el contrari). Seguint la teoria dels test no paramètrics basats en mesures ordinals, R_i calculat amb la mitjana sobre el conjunt dels problemes de prova ($j=1:N$) dóna una mesura sobre la bondat *global* de l'algorisme i:

$$R_i = \frac{1}{N}\sum_{j=1}^{N} R_{ij} \tag{5.4}$$

Amb aquesta definició, $R_i=1$ significaria que l'algorisme i seria sempre el millor, sobre qualsevol problema de prova, i per tant valor propers a 1 ens indiquen un bon comportament de l'algorisme en qüestió. A partir de la definició de R_i és possible calcular un valor per la distància crítica CD_α (veure l'apartat 7.4.2), que aporta informació sobre la diferència significativa entre els algorismes. A la figura 5.3, en què R es representa amb una altra magnitud de bondat X, es pot veure com aquesta informació també és analitzable gràficament: els algorismes amb uns valors de R més allunyats que $CD_{R,\alpha}$ respecte l'algorisme amb que es compara (en aquest cas, el de menor valor de R), es poden considerar significativament diferents en comportament d'aquest i, per tant, en aquests casos es podria rebutjar la corresponent hipòtesi nul·la.

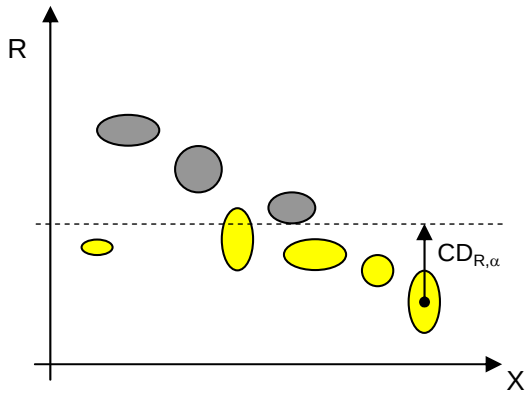

Figura 5.3: Exemple de representació per dues magnituds de bondat, R i X. Es representa també el valor de distància crítica calculat per R, que permet determinar quins són significativament diferents a l'algorisme amb menor valor de R, per al grau de significança α. Els algorismes representats per les el·lipses en groc no permetrien el rebuig de la hipòtesi nul·la respecte el de menor valor de R, mentre que les de color gris sí.

Aquest esquema gràfic aporta un guany molt important respecte les maneres habituals de representar els resultats obtinguts. Per veure-ho, es recupera el cas publicat a [2], en què el problema es situa en l'estudi de l'evolució del resultats obtinguts en un sistema CBR amb memòria de casos clusteritzada (SOMCBR, [21], en funció de l'estratègia de clusterització d'aquesta memòria.

En total, s'estudien 56 problemes de prova de dues classes, alguns d'ells provinents d'aplicacions mèdiques ([21]) i d'altres del repositori UCI ([3]). La complexitat dels problemes de prova és molt diversa, i això permet separarlos en un grup de baixa complexitat (17 problemes de prova), un grup amb complexitat moderada (un total de 30), i un grup menor d'elevada complexitat (9 problemes de prova). Les seves propietats es poden veure a la taula 5.1. Les diferents categories de complexitat han estat definides seguint allò expressat a l'apartat 4.7.

Per cada grup es vol analitzar el resultat d'aplicar fins a 12 estratègies diferents de clusterització (variants del SOMCBR), comparant-les sempre respecte el CBR estàndard, que actua com algorisme de control[5]. L'anàlisi cal fer-lo des de dos punts de vista: la precisió obtinguda en la classificació i

[5]Sempre i quan sigui el que millor comportament mostri des del punt de vista de la precisió. Com es veurà per als problemes de complexitat alta, això no té perquè ser sempre així.

Dataset	Atr.	Ins.	Tipus	Dataset	Atr.	Ins.	Tipus
segment2c2	19	2310	A	wav2c3	40	5000	B
iris2c2	4	150	A	wav2c1	40	5000	B
glass2c1	9	214	A	miasbi2c3	152	320	B
thy2c1	5	215	A	ddsm2c1	142	501	B
thy2c2	5	215	A	mias3c2c2	152	322	B
segment2c6	19	2310	A	thy2c3	5	215	B
segment2c7	19	2310	A	mias3c2c1	152	322	B
wine2c2	13	178	A	ddsm2c4	142	501	B
iris2c1	4	150	A	miasbi2c2	152	320	B
segment2c1	19	2310	A	wisconsin	9	699	B
wine2c1	13	178	A	wbcd	9	699	B
glass2c2	9	214	A	wav2c2	40	5000	B
miasbi2c4	152	320	A	sonar	60	208	B
glass2c4	9	214	A	wpbc	33	198	B
wine2c3	13	178	A	glass2c6	9	214	B
iris2c3	4	150	A	mias3c2c3	152	322	B
wdbc	30	569	A	biopsia	24	1027	B
segment2c3	19	2310	B	vehicle2c3	18	846	B
segment2c5	19	2310	B	vehicle2c2	18	846	B
glass2c3	9	214	B	bal2c3	4	625	C
vehicle2c1	18	846	B	bal2c2	4	625	C
segment2c4	19	2310	B	bal2c1	4	625	C
tao	2	1888	B	ddsm2c3	142	501	C
hepatitis	19	80	B	heartstatlog	13	270	C
glass2c5	9	214	B	μCa	21	216	C
ionosphere	34	351	B	ddsm2c2	142	501	C
vehicle2c4	18	846	B	pim	8	768	C
miasbi2c1	152	320	B	bpa	6	345	C

Taula 5.1: Descripció dels problemes de prova utilitzats per a l'assaig dels algorismes: nom, número d'atributs i instàncies, i tipus de complexitat segons s'ha definit a l'apartat 4.7. El sufix $2cX$ significa que el problema classifica la classe X respecte la resta de les classes, convertint així el problema de prova original en un de dos classes possibles. Els problemes de prova estan ordenats segons la regió de complexitat a la qual pertanyen.

el temps de càlcul en la fase de recuperació. Aquestes són les dues mesures de bondat a considerar en aquest problema, i l'objectiu és estudiar si alguna de les variacions del SOMCBR plantejades millora el comportament del CBR, en el sentit donat per alguna d'aquestes mesures de bondat.

De fet, per la pròpia construcció del SOMCBR, qualsevol de les estratè-

gies estudiades aportarà, en principi, un pitjor resultat en precisió respecte el CBR, doncs en tots ells els elements estudiats de la memòria de casos són menors que en el cas del CBR (veure [21] per la presentació completa d'aquesta metodologia). Ara bé, ni això és estrictament cert ([2]), ni la precisió és l'únic criteri a tenir en compte: cal també considerar el número d'elements de la memòria de casos que es fa servir en la fase de recuperació, en tant que d'aquesta magnitud en depèn fortament el temps d'execució.

L'aspecte de la precisió és estudiat a partir del rang definit per l'equació 5.4, mentre que la mida de la memòria de casos utilitzada ve donada per:

$$\%MC = log\left(\frac{\#}{\#_{CBR}}\right) \tag{5.5}$$

on $\#$ és el número d'elements de la memòria de casos utilitzats en l'estratègia en qüestió, i $\#_{CBR}$ són els utilitzats en el CBR estàndard, és a dir, la mida de tota la memòria de casos. Valors propers a 0 de $\%MC$ indiquen un ús molt elevat de la memòria, mentre que un valor negatiu de valor absolut elevat n'indica un poc ús i, per tant, un mètode computacionalment menys costós. El logaritme facilita la visualització de les diferències: per exemple, un valor de $\%MC = -1$ implica que s'ha utilitzat el 10% de la memòria de casos.

Seguint aquest plantejament, a la figura 5.4 es mostra l'anàlisi dels tres grups de problemes de prova, en funció de la seva complexitat (A equival a baixa complexitat, B moderada i C alta). També s'hi representa la distància crítica (CD_α), amb una barra vertical situada a partir de l'algorisme amb un menor valor de rang, que actua de control. D'acord amb la definició donada per CD_α, tots aquells algorismes situats dins la franja marcada per aquesta magnitud no tenen una diferència significativa dins l'interval de confiança determinat (en aquest cas $\alpha = 0.05$), pel que fa a la seva posició mesurada pel rang definit a l'equació 5.4, i respecte l'algorisme de millor resultat de rang, que s'utilitza com algorisme de control.

Com es pot observar en especial a la gràfica corresponent als problemes de prova d'alta complexitat, el cas del millor algorisme en precisió ($R_i \to 1$) no es correspon sempre amb el CBR. Ens trobem davant un exemple més de com l'anàlisi de resultats sense tenir en compte les propietats inherents d'aquests amagaria resultats importants, com el fet que el CBR no sigui el que aporti la millor precisió en alguns dels casos.

A banda del fet que la conclusió anterior difícilment es pot extreure si no és a partir d'aquests models gràfics, la valoració de la utilitat d'aquest plantejament gràfic es pot veure també en la comparació amb la taula 5.2, on es mostren les dades incloses a la figura 5.4, per bé que d'una manera

Prob. prova (pb)		**CBR**		**SOMCBR1**		**SOMCBR2**		...	**SOMCBR12**	
Nom	**Tipus**	**%Er.**	#	**%Er.**	#	**%Er.**	#		**%Er.**	#
Pb1	A	2.3	192	4.6	39	4.2	91	...	5.2	31
Pb2	A	4.2	192	4.7	104	4.9	147	...	6.0	32
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Pb18	B	16.8	924	22.6	168	20.5	453	...	23.6	98
Pb19	B	17.6	450	16.6	72	17.6	200	...	17.6	53
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Pb48	C	16.3	562	10.3	72	8.6	156	...	10.9	47
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Pb56	C	16.3	562	10.3	72	8.6	156	...	10.9	47

Taula 5.2: Taula d'exemple per analitzar el resultat de les 12 estratègies $\{SOMCBR1, ..., SOMCBR12\}$ respecte el CBR, aplicades sobre 56 problemes de prova $\{Pb1, ..., Pb56\}$. S'inclou la informació del percentatge d'error en la classificació ($\%Er.$) i del número d'operacions realitzades en l'etapa de recuperació (#). L'exemple intenta mostrar la magnitud que tindria una taula amb 13 algorismes sobre 56 problemes de prova, amb les dades estudiades a [2].

simplificada per qüestions d'espai. Malgrat els valors continguts a la taula continuen sent necessaris per una anàlisi més acurada dels resultats obtinguts (bàsicament, per aplicar-hi els test que es presentaran al capítol 7, a partir dels quals s'obté el valor de CD), és visible la utilitat de les gràfiques definides per a una primera interpretació de la bondat dels algorismes, i de quines són les comparacions que cal fer en un estudi més profund d'aquesta bondat.

5.3 Resum

En aquest capítol, s'han exposat en primer lloc els conceptes bàsics que permeten establir les hipòtesis a rebutjar o acceptar, a partir dels estadístics calculats amb els test d'inferència estadística que s'estudiaran en els capítols següents. S'ha justificat també el perquè de l'ús d'un nivell de significança $\alpha = 0.05$, la seva relació amb els errors "Tipus I" i "Tipus II", i s'han exposat els diferents conceptes que classifiquen les metodologies per a l'anàlisi dels resultats, i d'aquesta manera ha quedat justificat el plantejament que es farà en els capítols 6 i 7.

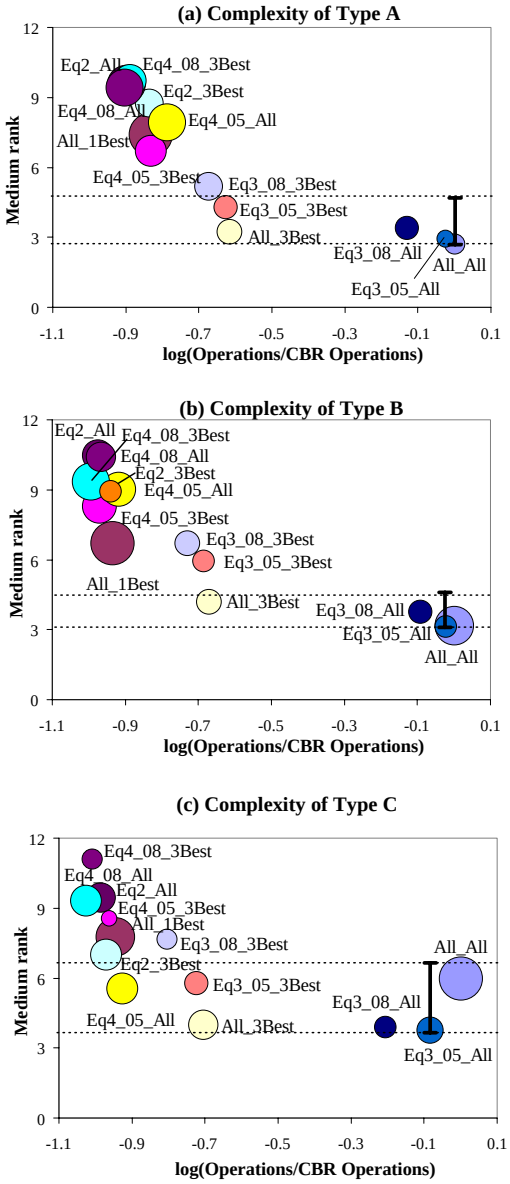

Figura 5.4: Anàlisi gràfica de les estratègies de recuperació en funció de la complexitat (A,B,C) dels problemes de prova. Cada circumferència representa els resultats d'un algorisme, amb el centre de la mateixa situada els valors mitjos definits a les equacions 5.4 i 5.5, i amb la seva àrea proporcional a la dispersió en la variable rang. S'hi observa com els resultats són clarament diferents depenent de la regió de complexitat (veure els casos com les configuracions *All_All* o *Eq4_05_All*). La nomenclatura utilitzada per a cada algorisme prové dels resultats publicats a [2], i es pot entendre a partir de l'esquema de la figura 9.1.

En la segona part, s'han desenvolupat diferents propostes per a la representació gràfica dels resultats obtinguts un cop assajats els algorismes sobre la col·lecció de problemes de prova. Especialment en el cas multivariant, les figures presentades permeten extreure unes primeres conclusions sobre els resultats, que estalvien en bona mesura la publicació de llargues taules numèriques de resultats. A més, el model de representació proposat aporta elements per decidir quin tipus de comparació es durà a terme, saber ràpidament si les propietats inherents dels problemes tenen un efecte important o no sobre les conclusions, i desenvolupar discusions més complertes sobre la bondat quan aquesta sigui mesurada per més d'una variable. Aquest és el cas, per exemple, dels resultats representats a la figura 5.4.

Capítol 6

Comparació simple de resultats

"A cynic might conclude that regardless of which method one employs,
there will always be reason to doubt the accuracy
of the probability value associated with the outcome of a study"
David J. Sheskin, [19]

En els capítols anteriors s'han posat les bases per a procedir a la comparació del comportament d'M algorismes, a partir dels resultats obtinguts sobre l'assaig en N problemes de prova. En el capítol 3 s'han introduït les mesures per establir la bondat d'un algorisme, i les maneres d'estimar-la, mentre que en el capítol 4 s'ha exposat la necessitat d'ampliar el camp d'estudi a aquelles propietats inherents als problemes de prova, sobre els quals s'assagen els M algorismes.

A partir d'aquí, én el capítol 5 s'ha iniciat la discussió sobre les eines per a efectuar aquestes comparacions. En primer lloc, s'ha introduït la nomenclatura habitual dels test d'hipòtesi i el seu significat, per procedir a continuació en l'anàlisi gràfica dels resultats obtinguts com a element de coneixement del problema, sovint en substitució de llargues taules de resultats.

El capítol que ara s'inicia estudia les metodologies per a la comparació del comportament de dos algorismes, que han estat assajats sobre la mateixa col·lecció de N problemes de prova. S'hi presenten les alternatives més habituals, l'aplicació de les quals suposa el compliment d'un conjunt de condicions sobre les dades obtingudes, que massa vegades són ignorades. La discussió es centra en el paper d'aquestes condicions que determinen el domini d'ús del test corresponent, i en els estudis que cal dur a terme per assegurar-ne el seu

compliment. El capítol continua amb una proposta de protocol d'aplicació de metodologies per a la comparació del comportament de dos algorismes, que serà verificada al capítol 8 a partir de la potència i la replicabilitat d'un test.

6.1 Plantejament

Existeixen multitud de metodologies per a comparar el comportament d'un conjunt d'algorismes, cadascuna aplicable a un tipus diferents de problema, d'acord amb les condicions pròpies d'aquest (número d'algorismes a comparar, número de problemes de prova, tipologia dels resultats obtinguts sobre cada un d'aquests problemes, etc.). Aquestes metodologies es poden agrupar segons diferents criteris: per la magnitud de la comparació i el número d'algorismes a comparar, per la tipologia de les dades que s'obtenen de l'aplicació dels algorismes sobre els problemes de prova (numèriques, ordinals, nominals), per les suposicions que es poden fer sobre aquestes dades, etc.

En els casos que s'estudien en aquest treball, les dades amb què es treballa sempre són numèriques i, si fa falta, es poden transfromar en ordinals o nominals. Per aquest motiu, i perquè la metodologia és extremadament diferent entre un cas i un altre, s'ha optat per introduir un primer nivell de separació en funció del número d'algorismes a comparar: en aquest capítol es tractaran només les metodologies per comparar els resultats de l'aplicació de dos algorismes, mentre que en el capítol següent es generalitzarà pel cas d'un número d'algorismes $M > 2$. La diferència de metodologies i el fet que aquest cas sigui habitual desaconsella fer el plantejament invers: primer l'estudi general per M algorismes, i després particularitzar per $M = 2$.[1]

Cal dir també que tots els exemples i càlculs que es presenten són "reals", és a dir, fruit de l'aplicació d'algorismes sobre problemes de prova realment utilitzats: es fugirà sempre de la utilització de dades sintètiques o preparades per a mostrar les mancances o virtuds d'algunes de les tècniques, com de vegades s'ha fet en alguns textos (vegi's, per exemple, un article molt referenciat de Dietterich, [13]). De fet, bona part dels resultats que es mostren han estat

[1] De fet, és bastant més habitual en casos "acadèmics" que reals: difícilment es publicarà una comunicació científica on un nou sistema classificador, per exemple, es compari només amb un altre classificador dels ja coneguts. Com es veurà al capítol 9, en la pràctica totalitat de casos estudiats s'involucraran més de dos algorismes. Tot i això, tractar en primer lloc el cas de la comparació simple facilita la comprensió de les metodologies per a comparacions múltiples, així com d'alguns dels errors més habituals que es cometen. Per això, es manté aquest ordre en el present treball.

ja publicats ([2], [4], [22], etc.).

Per aquesta raó, i com de fet passa sempre en la tipologia de problemes que s'estudiaran, les tècniques que s'analitzen són d'aplicació sobre mostrejos no independents, en el sentit que els algorismes s'apliquen sobre els mateixos conjunts de problemes de prova. En d'altres àmbits d'aplicació, com el mèdic, és habitual estudiar els resultats de tractament sobre conjunts independents de problemes de prova (és a dir, de pacients): no s'apliquen diferents tractaments sobre el mateix pacient, sinó un tractament diferent per a cada pacient. A banda, cal explicitar que els problemes utilitzats no es veuen afectats per l'aplicació d'un algorisme abans que un altre. Si per algun motiu això fos així, cap de les metodologies presentades serien utilitzables de la manera com en aquest treball s'exposen.

A partir d'aquí, el capítol comença amb l'anàlisi de les metodologies paramètriques (apartat 6.2), començant per les tècniques més simples i centrant-se ràpidament en l'estudi del t-test (apartats 6.2.2 a 6.2.4). D'ús molt habitual en l'àmbit que ens ocupa, és perfectament coneguda la forma d'utilitzar-lo, però no ho és tant el seu domini d'ús. És a dir, tot el conjunt de condicions que han de complir els problemes de prova i els resultats obtinguts de l'assaig dels algorismes per a què siguin vàlides les conclusions que d'ell se'n treuen.

Per aquest motiu, una de les aportacions d'aquest treball apareix al final d'aquests apartats, amb una proposta de protocol d'aplicació del t-test i l'estudi de les condicions que s'hi impliquen, per determinar quines són més determinants sobre el seu domini d'ús.

Aquest protocol és assajat sobre un exemple amb la comparació de tres algorismes, que permet estudiar les diferents situacions possibles. El protocol determina, en alguns casos, la necessitat d'utilitzar altres metodologies no-paramètriques. En l'apartat 6.3 s'exposen aquestes alternatives, tot centrant-se també en l'anàlisi del seu domini d'ús, i en les diferents capacitats d'analitzar les diferències significatives que puguin haver-hi.

Aquest apartat inclou dos exemples sobre dades ja publicades (a [2] i també a [4]). D'una banda, un mostra diferents situacions en les quals l'anàlisi incorrecte del domini d'ús de les metodologies implica unes conclusions errònies. L'altra exemple tracta una possible extrapolació dels tests presentats a un problema amb un número d'algorismes a comparar $M > 2$, tot presentant les conclusions parcials a què s'arriba amb una mala utilització de les metodologies presentades.

Més endavant, en el capítol 8, s'introduirà un canvi en l'enfocament del

que fins al moment s'ha discutit, que és interessant començar ja a entreveure, per la importància que donarà al citat protocol: de l'estudi de la bondat de l'algorisme es passarà a l'estudi de la bondat de la metodologia utilitzada. És a dir, quines variables poden donar una idea de com d'adequat és un determinat test d'inferència estadística per a l'estudi comparatiu de dos algorismes.

En aquest capítol, s'ha confiat la resposta a aquesta pregunta a les condicions que determinen el domini d'ús del test estudiat, junt amb el fet que un test paramètric sempre tindrà en compte més informació que un test no-paramètric. Per avaluar aquesta qüestió numèricament, i veure si les conclusions són coherents amb allò exposat en el domini d'ús, s'estudiaran en el capítol 8 dues mesures: la potència d'un test i la seva replicabilitat. Serà llavors quan quedarà ratificada la validesa del protocol proposat.

6.2 Test paramètrics

La comparació entre el comportament de dos algorismes és possible a partir dels resultats obtinguts de l'aplicació d'aquests sobre una col·lecció de problemes de prova. Les dades que s'obtenen es poden interpretar com a una mostra o estimació del comportament que aquests algorismes mostrarien si fossin assajats sobre tot l'univers de problemes existents: els resultats són, per tant, el resultat d'un mostreig de la distribució que determina els resultats de cada algorisme.

Els test paramètrics són aquells que, en el seu plantejament, pressuposen tot un conjunt de condicions sobre aquesta distribució, de la qual es coneix una mostra. Aquestes suposicions determinen un domini d'ús menor (hi haurà casos en què el no compliment d'aquestes condicions impedirà tenir garanties sobre les conclusions obtingudes) que, com a es veurà més endavant, vindrà compensat per una major capacitat de trobar aquelles diferències significatives existents entre els algorismes.

6.2.1 Mitjana sobre els problemes de prova

Una primera opció per a comparar el comportament de dos algorismes (A_1 i A_2) sobre un conjunt de problemes de prova és comparar el valor del valor mig de les mesures sobre tots els problemes de prova de què es disposa. És prou conegut que no es tracta d'una metodologia que sigui definitòria per treure'n conclusions ([39]), i tot just es pot considerar una dada de suport en les discussions que s'elaboren ([4]).

Els problemes més habituals amb aquesta magnitud són els relacionats amb els *outliers* i, en general, amb el fet que pocs valors allunyats de la mitjana podrien induir conclusions errònies sobre el comportament dels dos algorismes. En el citat treball [4], per exemple, diferències entre les mitjanes menors de l'1% amaguen comportaments significativament diferents, com es veurà més endavant. Fins i tot poden provocar conclusions errònies: un algorisme A_1 amb la mitjana de resultats menor que un altre A_2 pot ser millor, des del punt de vista de la significança estadística. Tan sols és necessari una certa presència de valors molt allunyats de la mitjana.

També es pot trobar un cas interessant de les enormes mancances de la mitjana sobre el conjunt de dades en el problema tractat a l'article [2], on s'estudia una metodologia general d'anàlisi de la fase de recuperació en un procés CBR amb la memòria de casos clusteritzada (SOMCBR): considerant els problemes de complexitat mitjana (veure l'apartat 5.2), apareix un *outlier* en la comparació dels mètodes OAN_05_NORM i OAN_08_NORM, provocat pel problema *wav2c1*. Si es considera aquesta dada, ambdós mètodes tenen una mitjana de comportament extremadament similar, mentre que si no es té en compte la diferència és ben visible: en l'apartat següent es mostrarà com aquesta diferència de comportament és estadísticament significativa, i per tant treballar només amb la mitjana sobre tots els problemes de prova no aportaria una conclusió correcta.

En resum, el càlcul de les mitjanes dels resultats sobre tots els problemes de prova pot aportar un primer nivell d'informació, però en cap cas és determinant per extreure'n cap conclusió, i fins i tot pot concloure erròniament.

6.2.2 Definició i càlcul del t-test

Estrictament parlant, el t-test és qualsevol test d'inferència estadística per a dos grups de mesures, l'estadístic del qual segueix una distribució t d'Student si la hipòtesi nul·la (H_0) és certa ([18]). Actualment, és una de les tècniques més conegudes i més utilitzades en l'anàlisi comparativa d'algorismes, si bé rarament s'utilitza tenint en compte totes les seves mancances i febleses, com es veurà en aquest apartat [2].

[2]Tot i el seu nom, poca gent coneix el fet que el seu autor és William Sealy Gosset, treballador de la destil·leria d'Arthur Guiness a Dublín. El senyor Guiness prohibia qualsevol tipus de publicació als seus treballadors, degut a una mala experiència anterior relacionada amb la divulgació d'informació confidencial de la seva empresa: si no fos per això, possiblement no parlaríem de la *distribució t d'Student*, sinó de la *distribució t de Gosset*.

En el t-test s'utilitzen les mesures de bondat dels dos algorismes aplicats sobre els N problemes de prova per determinar, a partir dels valors mitjans $\overline{X}_1$ i $\overline{X}_2$ obtinguts, si existeix una diferència real entre els valors de μ_1 i μ_2, mitjanes de la població de les bondats obtingudes de l'aplicació d'ambdós algorismes sobre tot l'univers de problemes de prova (quantitats que, òbviament, mai coneixerem). Aquest és el test apropiat en el cas que es desconeguin els valors reals de les variàncies del total de la població, com passarà sempre en els problemes estudiats. En cas contrari, caldria treballar amb el z-test de manera més apropiada ([93]): tot i això, és difícil imaginar l'entorn en què seria possible conèixer amb exactitud la variància d'una mesura sense conèixer-ne la seva distribució, o experimentalment el resultat sobre tots els problemes de prova existents.

Per tal d'aplicar-lo correctament, s'ha de tenir en compte que està basat en tres suposicions, dues de les quals són prou fortes ([19]), en el sentit que caldrà comprovar-les estrictament abans de donar credibilitat a la conclusió del test:

1. Aleatorietat de selecció: la mostra de N problemes de prova ha estat seleccionada a l'atzar d'entre la població de problemes de prova existents.

2. Normalitat en el resultat: la distribució de les dades obtingudes cal que segueixi una distribució de Gauss.

3. Homogeneitat de variàncies: la desviació dels resultats obtinguts sobre tota la població de problemes de prova existents cal que sigui igual per als dos algorismes testejats.

Si alguna d'aquestes tres condicions no es compleix, la fiabilitat del t-test es veu molt reduïda, i aleshores és recomanable utilitzar un altre test: estrictament parlant, no es podria aplicar. Aquí es pot veure el perquè, segons la definició donada anteriorment, aquest és un test paramètric, doncs per a la seva aplicació es necessita del compliment de tot un seguit de condicions.

La primera de les suposicions serà certa habitualment, en tant que es considera satisfeta si els problemes de prova no han estat triats per provocar un determinat resultat, sinó seguint els estàndards habituals: problemes del repositori UCI ([3]), problemes reals de domini mèdic, etc. En canvi, les altres dues suposicions caldrà que siguin comprovades sempre, cosa que es fa en molt poques ocasions.

Prova d'això s'obté amb un anàlisi simple que es pot fer sobre les publicacions presentades als principals congressos de l'àmbit del *machine learning* o

similar: per exemple, dels més de 100 articles acceptats al *23rd International Conference on Machine Learning* (ICML2006), en vuit d'ells s'utilitzava el t-test, i en cap d'aquests vuit es feia una comprovació explícita del compliment o no d'aquestes suposicions. És una costum habitual fer-ho d'aquesta manera, però no per això l'error és menor.

Per a realitzar el test sobre dos algorismes A_1 i A_2, cal obtenir l'estadístic t que es calcula com

$$t = \frac{\overline{D}}{s_{\overline{D}}} \tag{6.1}$$

En aquesta expressió, $\overline{D}$ és el valor mig de les diferències dels valors obtinguts per cada un dels N problemes de prova:

$$\overline{D} = \frac{1}{N}\sum_{j=1}^{N} D_j = \frac{1}{N}\sum_{j=1}^{N}(X_{1,j} - X_{2,j}) \tag{6.2}$$

i $s_{\overline{D}}$ representa l'error estàndard del valor mig de les diferències, calculat com

$$s_{\overline{D}} = \frac{\overline{s}_D}{\sqrt{N}} \tag{6.3}$$

on finalment $\overline{s}_D$ és l'estimació de la desviació estàndard sobre tota la població dels problemes de prova, que no es coneix però s'estima a partir de:

$$\overline{s}_D = \sqrt{\frac{\sum_{j=1}^{N} D_j^2 - \frac{(\sum_{j=1}^{N} D_j)^2}{N}}{N-1}} \tag{6.4}$$

Aquest valor de t es compara amb el valor crític $t_{crit,\alpha}$ que es pot trobar a les corresponents taules de la distribució t d'Student, donat un valor α de significança, i un cop determinat si l'anàlisi es fa per una o dues cues (direccional o no-direccional). La darrera magnitud necessària per conèixer $t_{crit,\alpha}$ són els graus de llibertat del problema, df, definits com:

$$df = N - 1 \tag{6.5}$$

En cas que el valor de t sigui superior a $t_{crit,\alpha}$, es considerarà que cal rebutjar la hipòtesi nul·la H_0 i que, per tant, el comportament dels dos algorismes és diferent a un nivell de significança determinat per α. Si no és així, no es tindran prous arguments per rebutjar H_0, i no es podrà considerar diferent el comportament dels dos algorismes a aquest nivell de significança.

De manera similar, la comparació amb el nivell de significança escollit es pot fer també a l'inrevés: l'estadístic t obtingut correspon a un valor p, que indica la probabilitat que la diferència observada sigui causada per una diferència real entre μ_1 i μ_2, i no per l'atzar. Si p és menor que l'α escollit, es diu que la diferència és significativa, cas que coincidirà amb un valor de $t > t_{crit,\alpha}$.

6.2.3 Domini d'úsdel t-test

La facilitat d'aquest càlcul, i el fet que la majoria de les eines de càlcul estadístic el facin sense gaire esforç computacional, ha extès possiblement l'ús en àmbits molt diferents de la recerca científica. A més, cal dir que el seu domini d'aplicació és prou ampli, i aquest fet el converteix en la primera opció en bona part de les contribucions que es publiquen. No obstant això, el domini d'aplicació o d'ús del t-test ve determinat bàsicament per tres factors no menyspreables, que sovint el delimiten més del que l'usuari suposa.

L'efecte de la presència d'*outliers*

El primer factor a tenir en compte el comparteix amb el que s'ha vist anteriorment per la mitjana sobre els problemes de prova: la presència d'algun *outlier* té un gran efecte sobre el càlcul de l'estadístic t, fins al punt que pot variar la conclusió sobre H_0 a la qual arriba.

Un bon exemple es troba recuperant els problemes de prova de complexitat mitjana utilitzats a l'article d'A.Fornells i altres abans comentat ([2]). Com es pot veure a la taula 6.1, la diferència en el percentatge d'error obtingut (mesura que s'utilitza com determinant de la bondat dels algorismes) pel problema $wav2c1$ en la comparació dels mètodes A_1 (que correspon a l'algorisme OAN_05_NORM) i A_2 (OAN_08_NORM) és molt superior a la mitjana, suficient com per considerar-lo un *outlier*.

Aquest valor provoca el que es resumeix a la taula 6.2: la conclusió obtinguda per l'anàlisi a partir del t-test és ben diferent, depenent de si es considera o no el valor obtingut pel problema de prova $wav2c1$. Considerant aquesta dada el dos algorismes tindrien un comportament significativament equivalent, mentre que al eliminar-la del conjunt de dades a considerar s'obté la conclusió oposada, amb uns valors de t que permeten rebutjar H_0 amb un nivell de significança $\alpha = 0.05$.

Una primera conclusió, per tant, és que la presència d'*outliers* delimita

Dataset	X_1	X_2	$X_1 - X_2$	**Dataset**	X_1	X_2	$X_1 - X_2$
biopsia	23.54	23.56	-0.02	segment2c4	3.67	2.99	0.68
ddsm2c1	17.22	17.61	-0.39	segment2c5	6.23	7.12	-0.89
ddsm2c4	25.2	23.95	1.25	sonar	31.37	33.29	-1.92
glass2c3	9.7	10.4	-0.7	tao	6.25	7.2	-0.95
glass2c5	26.64	25.94	0.7	thy2c3	7.79	10.47	-2.68
glass2c6	28.74	29.2	-0.46	vehicle2c1	13.65	14.01	-0.36
hepatitis	23.88	24.19	-0.31	vehicle2c2	27.39	27.78	-0.39
ionosphere	18.16	19.51	-1.35	vehicle2c3	29.31	27.9	1.41
mias3c2c1	27.87	26.24	1.63	vehicle2c4	11.38	13.51	-2.13
mias3c2c2	15.53	15.06	0.47	wav2c1	18.64	4.86	13.78
mias3c2c3	32.92	33	-0.08	wav2c2	21.59	21.86	-0.27
miasbi2c1	17.19	17.74	-0.55	wav2c3	18.37	18.7	-0.33
miasbi2c2	27.58	28.28	-0.7	wbcd	4.68	5.22	-0.54
miasbi2c3	22.11	22.89	-0.78	wisconsin	4.76	5.44	-0.68
segment2c3	4.38	4.64	-0.26	wpbc	26.64	28.91	-2.27

Taula 6.1: Percentatge d'error en l'aplicació de les estratègies sobre els problemes de prova de complexitat mitjana, OAN_05_NORM (X_1) i OAN_08_NORM (X_2). La darrera columna conté el valor de la diferència entre aquests dos resultats, on es pot observar la diferència del que s'obté per $wav2c1$ respecte la resta.

	Amb wav2c1	**Sense wav2c1**
Mitjana de la diferència	0.03	-0.44
Desviació estàndard de la diferència	2.79	1.02
Número de problemes de prova	30	29
Estadístic t	0.06	2.33
Valor de p	0.95	0.03

Taula 6.2: Resum de l'anàlisi sobre les dades de la taula anterior. Es pot observar com la presència de les dades corresponents al problema de prova $wav2c1$ porta a la conclusió que els dos algorismes es comporten de manera equivalent ($p >> 0.05$), mentre que quan no es considera aquesta dada el comportament indica dos algorismes diferents ($p < 0.05$), perquè es pot rebutjar H_0.

en gran mesura el domini d'aplicabilitat del t-test. Davant d'aquesta situació les habituals estratègies ([94]) passen per eliminar un determinat percentatge de les mostres (*trimming*), substituir aquelles extremes per les més properes (*winsorizing*) o aplicar certes transformacions a les dades per eliminar l'efecte

d'aquests punts (arrels, logaritmes,...). De totes maneres, donat el caràcter de les discussions que ens ocupen, possiblement el millor és eliminar les n dades que mostrin aquest caràcter de gran desviació respecte la mitjana, determinar el caràcter dels algorismes sobre els $N - n$ problemes restants i, en tot cas, analitzar a banda el comportament en els n problemes que provoquen els *outliers*, sempre i quan es compleixi que $n << N$.

La normalitat dels resultats obtinguts

El segon element a tenir en compte és la suposició que es fa sobre la normalitat de les dades que s'obtenen. És a dir, per poder aplicar el t-test correctament, la diferència $X_1 - X_2$ hauria de seguir una distribució gaussiana. Per tant, en el cas de testejar-ho sobre N problemes de prova, cal que les N dades $(X_{1,j} - X_{2,j})$ obtingudes segueixin un comportament que pugui provenir d'una distribució normal. Per valors de N elevats no acostuma a haver-hi problema (tinguem present tota la teoria al voltant del teorema central del límit, [95]), però la realitat és que sovint tenim un conjunt de problemes de prova petit, ja sigui degut a l'especificitat del domini o bé al cost computacional d'assajar els algorismes sobre grans quantitats de problemes.

En cas que ens trobem amb aquesta realitat, cal aplicar alguns dels test existents per comprovar la normalitat de les dades, i veure així si es complexi la segona de les suposicions necessàries per aplicar el t-test. La base de tots aquests tests és la prova de Kolmogorov-Smirnov (test K-S, [96]), que compara la funció de distribució F_N que es pot construir empíricament a partir de les observacions y_j

$$F_N(x) = \frac{1}{N} \sum_{j=1}^{N} \begin{cases} 1 \text{ si } y_j < x \\ 0 \text{ altres} \end{cases} \tag{6.6}$$

amb la funció de distribució de comparació F (en el nostre cas la distribució normal). La comparació es du a terme amb dos estadístics que donen una mesura de probabilitat per la similitud d'ambdues distribucions, i que es calculen de la següent manera:

$$D_N^+ = max(F_N(x) - F(x)) \tag{6.7}$$

$$D_N^- = max(F(x) - F_N(x))$$

Igual que en els altres casos, per valors inferiors a 0.05 considerarem rebutjar aquesta similitud. També en aquest cas, l'efecte dels *outliers* és molt

important, com es pot veure en l'anàlisi fet a partir de les dades mostrades a la taula 6.1, les gràfiques de la qual apareixen a la figura 6.1. En aquesta figura es representen els histogrames de les dades obtingudes per les diferències $X_{1,j} - X_{2,j}$ respecte la corva normal equivalent (en mitjana i variança): en el segon cas s'observa una major dificultat per comprovar la similitud amb la distribució normal.

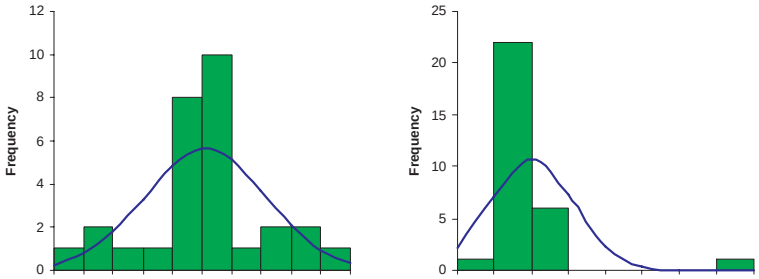

Figura 6.1: Histogrames de la diferència dels resultats pels algorismes discutits. En el primer cas (OAN_05_NORM), no s'hi inclou el valor sobre el problema de prova $wav2c1$, que sí apareix en el segon (OAN_08_NORM).

Una variació a aquest test és el conegut com a test de Lilliefors ([97]), que no especifica a priori els valors de mitjana i variança de la distribució. A partir d'aquí, diverses estratègies han estat desenvolupades per millorar petites desviacions trobades en el test de K-S, i no hi ha un clar consens sobre quina és la més fiable. Força autors (veure, per exemple, [98] o [99]) prefereixen el test de Jarque-Bera ([100]), que desenvolupa un estadístic $J-B$ de comparació amb la distribució normal

$$JB = \frac{N}{6}\left(S^2 + \frac{(K-3)^2}{4}\right) \tag{6.8}$$

a partir dels valors de la desviació (S) i la curtosis (K), definides com

$$S = \frac{\mu_3}{(\sigma^2)^{3/2}} \tag{6.9}$$

$$K = \frac{\mu_4}{(\sigma^2)^2}$$

amb μ_3 i μ_4 com a moments centrats de tercer i quart ordre, respectivament ([95]). L'estadístic JB tendeix a una distribució chi-quadrat amb dos graus de llibertat si el comportament de les dades s'acosta a una gaussiana, i això permet calcular una probabilitat de rebuig de la hipòtesi de semblança a aquesta distribució.

Finalment, el test de Shapiro-Wilk ([101]) i el test d'Anderson-Darling ([102]) són considerats també per altres autors ([103]) com una bona alternativa per detectar problemes amb la normalitat de les dades. Ambdós generen un estadístic que, a través de la comparació amb el corresponent valor crític, determina la conveniència de rebutjar la hipòtesi nul·la pel grau de confidència determinat.

A la taula 6.3 s'observa el resum dels resultats obtinguts per tots aquests test aplicats sobre les dades mostrades anteriorment. En primer lloc cal destacar que, malgrat la gran diferència entre els valors obtinguts, les conclusions a què arriben tots els test són equivalents: permeten rebutjar la hipòtesi de normalitat amb un nivell de significança estadística de 0.05 si es considera el valor *outlier*, i no ho permeten si no se'l considera. En segon lloc, cal destacar el resultat en si mateix, on es veu el gran efecte de no fer un bon tractament sobre els valors llunyans de la mitjana. Finalment, les diferències entre els resultats que es troben en el cas que no es consideri l'*outlier* fan necessari un estudi més acurat per cada test en funció de les característiques del conjunt de problemes de prova que s'analitzin.

Test	**Amb wav2c1**		**Sense wav2c1**	
	Estadístic	**p-value**	**Estadístic**	**p-value**
K-S	1.54	< 0.01	0.85	0.08
J-B	468.50	< 0.01	0.04	0.98
S-W	0.51	< 0.01	0.95	0.19
A-D	4.51	< 0.01	0.73	0.06

Taula 6.3: Mesures de normalitat, segons els 4 test definits al text. S'observa com, malgrat la diferència de valors obtinguts per als estadístics i el propi valor de p, les conclusions són coherents en tots els casos.

Per totes aquestes mesures de normalitat s'ha de tenir present que la seva capacitat es veu molt reduïda si el número de dades (és a dir, de problemes de prova) de què es disposa és petit. De fet, sovint és tan petit que és pràcticament impossible que aquests test siguin capaços de discriminar un comportament allunyat de la normalitat estadística, en cas que existeixi. Per sort pel nostre plantejament, l'experimentació ens mostrarà posteriorment com no és aquesta la condició més determinant sobre el domini d'aplicabilitat del t-test.

L'homogeneïtat de les variàncies

El darrer element a tenir en compte afecta la tercera suposició pel t-test, coneguda com a l'homogeneïtat de les variàncies: la desviació dels resultats obtinguts sobre tota la població de problemes de prova existents cal que sigui igual per als dos algorismes testejats. En tant que els algorismes s'apliquen sobre el mateix conjunt de problemes de prova, no és apropiat utilitzar el test de F_{max} ([19]), que sí valdria si els problemes de prova fossin diferents per cada algorisme.

En el nostre cas, cal tenir en compte el valor diferent de zero que pren la correlació entre els resultats dels dos algorismes aplicats, definida com:

$$\rho_{X_1,X_2} = \frac{Cov(X_1, X_2)}{\sigma_1 \sigma_2} \tag{6.10}$$

Tenint en compte aquest fet, es defineix l'estadístic t per al test sobre l'homogeneitat de variàncies de la següent manera:

$$t = \frac{(\overline{s}_M^2 - \overline{s}_m^2)^2 \sqrt{N-2}}{\sqrt{4\overline{s}_M^2 \overline{s}_m^2 (1 - \rho_{X_1,X_2})}} \tag{6.11}$$

on $\overline{s}_M^2$ i $\overline{s}_m^2$ són els valors màxim i mínim, respectivament, d'entre els dos valors de $\overline{s}$ estimats pels dos algorismes. Aquest estadístic segueix una distribució d'Student amb graus de llibertat iguals a $N-2$ i, per tant, amb l'habitual manera de fer es pot comparar amb el corresponent valor crític $t_{crit,\alpha}$ i extreure'n el valor p, que permet decidir si es rebutja la hipòtesi d'igualtat de variàncies pel nivell de confidència determinat. Altres opcions es poden consultar a [104] o [105].

Un bon exemple per aquest test es troba en la comparació de les estratègies EBN_MAX_3, $PEBN$ i $PEBN\_MAX\_3$, definides a l'article de Fornells i altres abans citat com a variants del SOMCBR ([2]), per als problemes de prova d'alta complexitat. Els resultats obtinguts després de l'aplicació dels algorismes es mostren a la taula 6.4 (la mesura de la bondat aquí ve donada per l'error de classificació en la fase de test) i el resum pel que fa a l'homogeneïtat de les variàncies es pot veure a la taula 6.5.

En aquests casos, els resultats mostren com el rebuig de la hipòtesi d'homogeneïtat de variàncies invalida la conclusió a què ens portaria el t-test. En el primer cas (comparació entre EBN_MAX_3 i $PEBN$, X_1 vs X_2), el t-test conclou rebutjar H_0 a un nivell de significança $\alpha = 0.05$, doncs el valor de p obtingut és menor, però l'estudi de les variàncies mostra com, al mateix nivell de significança, no es compleix la tercera de les condicions per a

Dataset	X_1	X_2	$X3$
bal2c1	8.56	9.12	9.36
bal2c2	13.36	12	11.56
bal2c3	10.56	11.92	11.44
bpa	37.39	44.93	43.04
ddsm2c2	34.43	37.27	36.78
ddsm2c3	34.78	36.23	36.63
heart-statlog	23.51	26.11	25.56
mamografies	37.15	37.73	35.76
pim	28.62	32.39	31.12

Taula 6.4: Percentatge d'error en l'aplicació de les estratègies sobre els datasets de complexitat alta (EBN_MAX_3 (X_1), $PEBN$ (X_2) i $PEBN\_MAX\_3$ (X_3)).

	X_1 **vs** X_2	X_1 **vs** X_3	X_2 **vs** X_3
p-value (t-test)	0.034	0.089	0.033
Rebuig H_0 ($\alpha = 0.05$)	Si	No	Si
$\overline{s_M}$	13.35	12.91	13.35
$\overline{s_m}$	11.8	11.8	12.91
N	9	9	9
ρ_{X_i,X_j}	0.988	0.998	0.998
df	7	7	7
t-test (hom. var.)	2.94	6.44	2.35
$t_{crit}(.05)$	2.37	2.37	2.37
$t_{crit}(.01)$	3.5	3.5	3.5
p-value (hom. var.)	0.022	$<$0.01	0.05
Rebuig hom. var. ($\alpha = 0.05$)	Si	Si	No

Taula 6.5: Resum dels resultats per als algorismes EBN_MAX_3 (X_1), $PEBN$ (X_2) i $PEBN\_MAX\_3$ (X_3), pel que fa al t-test (amb les corresponents conclusions sobre el rebuig de la hipòtesi nul·la), i pel que fa al compliment de l'homogeneïtat de variàncies, que es pot rebutjar en els dos primers casos.

l'aplicació del t-test: per tant, no es pot considerar vàlida la conclusió sobre H_0 a la qual s'arriba.

El segon cas (EBN_MAX_3 i $PEBN\_MAX\_3$, X_1 vs X_3) és potser el més rellevant: no apareix una diferència significativa (doncs $p > 0.05$), però la diferència entre els variàncies impedeix considerar correcta aquesta conclusió.

De fet, les dades d'error ens mostren com l'algorisme EBN_MAX_3 (X_1) es comporta millor en 7 de les 9 ocasions al $PEBN\_MAX\_3$ (X_3), i això fa pensar que potser algun mètode no-paramètric hi mostrarà una diferència significativa.

En el darrer cas (X_2 vs X_3), la conclusió del t-test (diferència significativa doncs $p < 0.05$) pot ser considerada vàlida a nivell de la condició sobre l'homogeneïtat de les variàncies, doncs l'estadístic t obtingut en l'homogeneïtat de variàncies és menor que el nivel crític per $\alpha = 0.05$: es rebutja la hipòtesi nul·la que enuncia igualtat de comportament entre ambdós algorismes, però no es rebutja la que enuncia homogeneïtat de variàncies.

Sobre aquest mateix exemple, cal preguntar-se si el rebuig de la hipòtesi sobre les variàncies esdevé quelcom comú en les comparacions per parelles que es fan. El cas utilitzat d'exemple té una particularitat especial: són els problemes de prova dels quals es coneix que la complexitat, mesurada d'acord a l'article d'A.Fornells i altres on es defineixen les regions de complexitat ([22]), és més elevada. Això fa que l'error comès per l'algorisme també sigui, en general, major, i això possiblement facilita que els valors de les dispersions siguin elevats. Si això no passa, i per tant es treballa amb problemes de prova de baixa complexitat, difícilment els valors obtinguts per les variàncies no compliran la condició per aplicar el t-test.

Dit d'una altra manera, l'estudi de l'homogeneïtat de les variàncies, en les variants del SOMCBR estudiades, esdevé crític en aquells problemes de prova amb una complexitat més elevada, i aquest fet mostra novament la importància d'allò introduit al capítol 4: l'estudi de les propietats inherents dels problemes de prova és fonamental per a la correcta interpretació dels resultats que s'obtindran posteriorment.

Sigui com sigui, sembla evident la necessitat de:

1. Establir un protocol ben robust sobre les condicions per aplicar el t-test, que asseguri que el problema estudiat està dins el domini d'aplicabilitat d'aquesta tècnica d'inferència estadística.

2. Conèixer alternatives a l'aplicació del t-test, des d'un punt de vista dels tests no-paramètrics.

Aquest segon punt serà analitzat en profunditat a l'apartat 6.3, mentre que l'esquema que es proposa a continuació és una bona guia per a l'estudi del domini d'aplicabilitat del t-test, i per tant ve a respondre a la qüestió plantejada. Tenint en compte el comentari realitzat anteriorment sobre els article del darrer *23rd International Conference on Machine Learning*

(ICML2006), i altres exemples que es podrien trobar, sembla especialment interessant aplicar el que es proposa a la figura 6.2, a mode de protocol per a la comprovació del domini d'ús del t-test. Posteriorment, en el següent apartat, s'analitzarà quina de les condicions és més crítica per a establir correctament el domini d'aplicabilitat del t-test, i si serà permès relaxar les condicions en alguns casos.

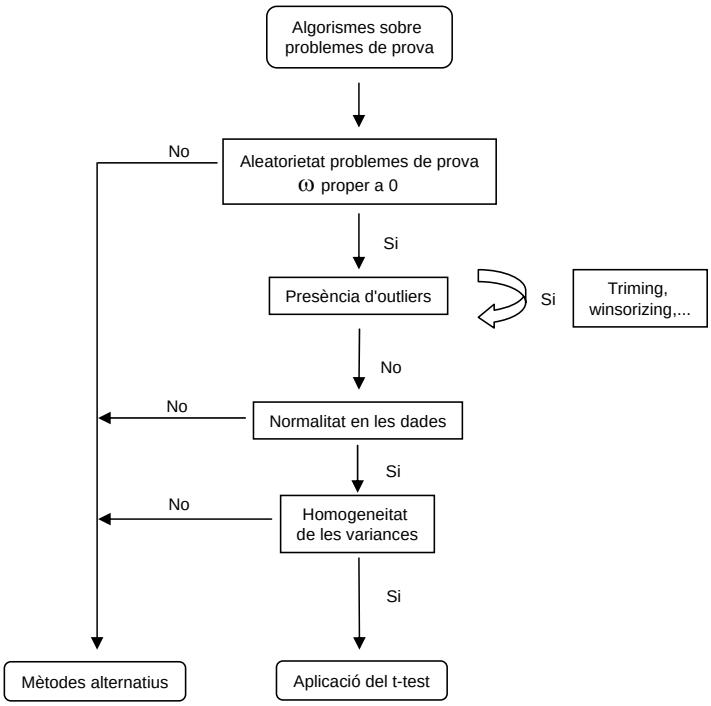

Figura 6.2: Proposta de protocol per a la correcta aplicació del t-test en la comparació de resultats de dos algorismes sobre un conjunt d'N problemes de prova.

6.2.4 Altres aspectes sobre el t-test

A banda d'allò comentat fins ara, es poden tenir presents altres consideracions "menors" sobre aquesta metodologia, en el sentit que habitualment no seran d'aplicació o utilitat en els casos que s'estudien. No obstant això, de cara a la completesa del treball, s'hi dedica una breu atenció.

D'una banda, habitualment es parla de la possibilitat d'avaluar la potència d'un mètode, entesa com la capacitat de rebutjar la hipòtesi nul·la H_0 si

és certa la hipòtesi alternativa, H_1. Aquesta capacitat és complementària a la probabilitat de cometre un error "Tipus II", que creix quan es redueix la probabilitat de cometre un error "Tipus I" (és a dir, quan es redueix el llindar de significació estadística). Per tant, es pot deduir que si es manté aquest llindar en valors moderats (com l'habitual $\alpha = 0.05$ que s'utilitza) l'error de "Tipus II" es mantindrà també en nivells baixos, i això dotarà d'una major capacitat o potència el mètode en qüestió.

Dit d'una altra manera: la reducció del valor de α, per voler augmentar la seguretat en una conclusió de rebuig d'H_0, porta a l'augment de la probabilitat de cometre un error de "Tipus II" i, per tant, redueix la capacitat de detectar diferències significatives quan aquestes existeixin. El raonament és cert també a la inversa. Per tant, si per algun motiu es considera que el llindar de significança marcat no permet concloure significativament sobre unes diferències que es sospita que existeixen, l'alternativa més recomanable és augmentar el número N de problemes de prova sobre els quals s'assagen els algorismes, no modificar en gran quantitat el valor d'α.

Més endavant, en analitzar la potència i la replicabilitat d'un test de comparació (capítol 8), es discutirà més profundament la relació entre la probabilitat de cometre ambdós tipus d'errors, i la relació d'això amb el domini d'aplicabilitat del test analitzat.

D'altra banda, es descriu com calcular el que es coneix com la magnitud de l'efecte del "tractament", en el nostre cas l'efecte dels algorismes utilitzats sobre els problemes de prova. El fet és que les condicions d'aplicació dels algorismes corresponents poden estar subjectes a una certa variabilitat (molt més evident en casos de tractaments-pacients, en què el tractament pot dependre de molts imponderables), i aquesta variabilitat pot afectar els resultats obtinguts sobre els problemes de prova. Si no es coneix en quin percentatge es poden veure afectats, com es pot descartar la possibilitat que el resultat del t-test vingui fortament condicionat per aquesta variació induïda per les pròpies condicions d'assaig de l'algorisme?

En tot cas, en els casos que ens ocupen aquesta variabilitat hauria de ser mínima, doncs el control sobre les condicions d'aplicació dels algorismes és pràcticament total. Aquest efecte es pot mesurar através de l'estadístic ω^2 ([106]), definit com:

$$\omega^2 = \frac{SS_{BC} - MS_{Res}}{SS_T + MS_{Res}} \tag{6.12}$$

on les quantitats SS_{BC}, MS_{Res} i SS_T es definiran posteriorment en l'apartat 7.3.1 (la referència permet estalviar-se una definició parcial i fora de context en aquest punt). Segons Cohen ([107]), valors de $\omega^2 < 0.06$ garanteixen

uns efectes prou petits de les condicions d'aplicació dels algorismes com per no tenir cap tipus d'influència en el resultat dels test de comparació. En tots els casos que es mostraran, els valors d'ω^2 difícilment superaran valors de 0.01, com es comprovarà en l'apartat següent, i per tant és cert que, en la tipologia de problemes tractada en aquest treball, aquests factors seran extremadament menors.

6.2.5 Un exemple: problemes de prova de dominis mèdics

Per tal de mostrar clarament les mancances d'un anàlisi per t-test sense consideracions prèvies, i la necessitat d'establir alternatives no-paramètriques en alguns casos, es finalitza l'apartat amb la presentació d'un exemple en què els problemes de prova pertanyen a un mateix àmbit: considerem el subconjunt de 18 problemes de prova que pertanyen al domini de problemes mèdics, d'entre els utilitzats a l'article de Fornells i altres recurrentment referit en aquest apartat ([2]). A la taula 6.6 es poden veure els resultats, expressats a partir de l'error de classificació en la fase de test per cadascun dels tres algorismes escollits (tots ells variacions del SOMCBR, definits també a la citada publicació).

En les tres comparacions per parelles possibles s'observen tres comportaments ben diferents, amb un denominador comú: en cap cas es compleixen totes les condicions com per aplicar el t-test, i per tant en principi calen alternatives no-paramètriques. Els resultats es poden veure resumits a la taula 6.7.

En el segon cas es comparen els resultats per X_1 i X_3, valors de l'error corresponent als algorismes OAN_08 i OAN_05_NORM. Les dades resultants de les diferències pel conjunt dels problemes de prova no compleixen la condició de normalitat, i a més violen l'homogeneïtat de variàncies, en ambdós casos a un nivell $\alpha = 0.05$ de significança estadística. El mateix passa per la comparació entre $OAN\_05\_MAX\_3\_NORM$ (X_2) i OAN_05_NORM (X_3): per tant, en cap dels dos casos es pot aplicar el t-test amb garanties de fiabilitat per la conclusió. Els resultats trobat, que indica que en cap d'aquests dos casos el t-test donaria arguments per a rebutjar la condició nul·la H_0 ($p > 0.05$), no poden ser tinguts en compte.

En canvi, en el primer cas es comparen els algorismes OAN_08 (X_1) amb $OAN\_05\_MAX\_3\_NORM$ (X_2), i allí no es viola la condició d'homogeneïtat de variàncies. No obstant això, les dades segueixen sense complir les condicions de normalitat necessàries, que a priori permetrien aplicar el t-test (implicant el rebuig de la hipòtesi nul·la, doncs $p < 0.05$).

Dataset	X_1	X_2	$X3$
biopsia	18.62	22.76	23.54
ddsm2c1	17.32	17.42	17.22
ddsm2c2	35.43	36.33	27.65
ddsm2c3	34.38	36.42	36.58
ddsm2c4	22.86	23.65	25.20
heart-statlog	22.78	24.82	24.63
hepatitis	29.68	23.07	23.88
mamografies	37.39	39.12	36.46
mias3c2c1	22.21	26.01	27.87
mias3c2c2	12.58	13.90	15.53
mias3c2c3	29.12	34.78	32.92
miasbi2c1	13.51	18.91	17.19
miasbi2c2	24.61	29.14	27.58
miasbi2c3	19.53	23.36	22.11
miasbi2c4	11.41	13.20	14.06
wbcd	4.40	5.15	4.68
wdbc	4.22	5.98	5.93
wisconsin	4.12	4.47	4.76

Taula 6.6: Percentatge d'error en l'aplicació de les estratègies OAN_08 (X_1), $OAN\_05\_MAX\_3\_NORM$ (X_2) i OAN_05_NORM (X_3) sobre problemes de prova de a dominis mèdics.

Aquest comportament tant extrem ve donat, en part, pel que ja s'ha comentat dels problemes d'elevada complexitat: en tant que problemes de prova reals i pertanyents al domini mèdic, habitualment incorporen força elements de complexitat i, de fet, només 2 dels 18 poden ser considerats amb complexitat "baixa". Això provoca valors considerables per l'error i les seves variàncies, i faciliten la violació de les condicions per l'aplicació del t-test, especialment en el cas de les condicions de normalitat.

Ara bé, es pot fer una interpretació menys restrictiva d'aquestes condicions, per poder establir un domini d'ús del t-tes més ampli? En cas contrari, el que s'acaba d'argumentar descartaria pràcticament per complert l'ús del t-test en col·leccions de problemes d'elevada complexitat, i ho deixaria en dubte per la resta de casos. En l'apartat següent, es presentaran les alternatives no-paramètriques que existeixen, el seu domini d'ús i propietats, i la seva capacitat per determinar les diferències significatives que puguin existir.

	X_1 **vs** X_2	X_1 **vs** X_3	X_2 **vs** X_3
p-value (t-test)	0.01	0.12	0.31
Rebuig H_0 ($\alpha = 0.05$)	Si	No	No
p-value (S-W)	<0.01	<0.01	<0.01
Normalitat	No	No	No
N	18	18	18
t-test (hom. var.)	0.31	2.62	2.92
$t_{crit}(.05)$	2.37	2.37	2.37
$t_{crit}(.01)$	3.5	3,5	3,5
Rebuig hom. var. ($\alpha = 0.05$)	No	Si	Si
ω^2	<0.01	0.01	<0.01

Taula 6.7: Resum dels resultats per als algorismes comparats. El resultat X_1 correspon al de l'agorisme OAN_08, X_2 al de l'algorisme $OAN\_05\_MAX\_3\_NORM$, i X_3 al de OAN_05_NORM.

Amb les mateixes dades que s'acaben d'analitzar, es veurà com les restriccions fins aquí exposades sobre el t-test es poden relaxar, i que el domini d'ús d'aquest pot ser parcialment ampliat, en funció dels propis valors de l'estadístic t i de $t_{crit,\alpha}$.

6.3 Alternatives no paramètriques

L'exemple amb què acaba la secció anterior posa de manifest la necessitat d'establir alternatives per a l'anàlisi comparatiu del comportament de dos algorismes, quan les condicions per aplicar els test paramètrics (en aquest cas el t-test) no es compleixen. Així doncs, cal establir metodologies no-paramètriques, tal i com han estat definides a l'apartat 5.1.3.

6.3.1 Introducció als test no-paramètrics

Les metodologies no-paramètriques són en general bastant poc utilitzades en l'àmbit que ens ocupa, malgrat estar profundament treballades des d'un punt de vista teòric. Demsar ([9]) va estudiar la manera com s'obtenien i comparaven els resultats en els prop de 500 articles acceptats en les edicions del 1999 al 2003 de l'*International Conference on Machine Learning* (ICML),

dels quals 174 aplicaven tècniques que són objecte de discusió en aquest treball. Els casos en què els autors aplicaven tests no-paramètrics no superaven el 15% d'aquests articles, i pràcticament sempre eren un simple contatge de guanys i pèrdues, que com es veurà no és la tècnica més recomanable per discriminar diferències de comportament entre els algorismes. En l'edició del 2006 d'aquest mateix Congrés, per exemple, ni un sol dels autors parlava del test de Wilkonson, el principal enfocament no-paramètric d'aquests problemes.

De la mateixa manera que no té sentit ignorar aquestes metodologies (tinguem present que si no s'estudien les condicions per aplicar el t-test difícilment cal pensar en una alternativa), tampoc són certes opinions exposades en alguns articles recents (vegi's el mateix Demsar, [9]), segons les quals l'esforç per estudiar les condicions d'aplicabilitat dels test paramètrics, i la dificultat que es compleixin aquestes condicions, fa recomanable utilitzar sempre test no-paramètrics. Com es veurà més endavant, en estudiar la potència i la replicabilitat dels tests, les condicions sobre el t-test es poden relaxar i, en cas que sigui possible confiar en les seves conclusions, la seva capacitat per discriminar diferències significatives sempre és major. Precisament per aquesta major capacitat, val molt la pena l'estudi del domini d'ús d'un test paramètric.

Les diferències principals entre els tests no-paramètrics i els paramètrics son dues: les suposicions prèvies i les maneres com es tracten els resultats obtinguts de l'assaig dels dos algorismes sobre els problemes de prova. Pel que fa a la primera, s'entén per test no-paramètric aquell que no efectua cap suposició sobre la distribució del resultat que s'observa per a cada algorisme i, per tant, es pot aplicar sempre. Com veurem, això és cert sempre que es compleixi la condició d'aleatorietat de selecció dels problemes de prova.

Referent a les dades que s'obtenen, i a diferència dels test paramètrics (en què es treballa amb el resultat numèric dels algorismes sobre els problemes de prova), els tests no-paramètrics acostumen a transformar aquests resultats en una relació d'ordres o rangs, per establir quins problemes de prova són aquells que provoquen major diferència entre ambdós algorismes, i el signe d'aquesta diferència. Aquesta operació fa que es perdi informació sobre els resultats, però es guanyi en domini d'ús de la metodologia, perquè no cal assumir massa condicions sobre les dades per poder aplicar-los.

6.3.2 Test de signes de Wilcoxon

El test de signes de Wilcoxon (en anglès referit sovint com *Wilcoxon signed-ranks test*) és la principal alternativa al t-test, quan no es compleixen les condicions per a l'aplicació d'aquell. El principal avantatge és que es basa en condicions molt més febles que l'anterior. Bàsicament, cal que la mostra de N problemes de prova hagi estat seleccionada a l'atzar d'entre la població de problemes de prova existents, i que els resultats estiguin en un format que permetin ordenar-los. Són realment condicions molt poc restrictives.

L'estadístic es contrueix de la manera següent. Es defineix en primer lloc d_j com la diferència entre els valors obtinguts per cada algorisme sobre el problema de prova j:

$$D_j = X_{1,j} - X_{2,j} \tag{6.13}$$

Aquests valors són ordenats d'acord amb el seu valor absolut, i a cada problema de prova se li atorga un rang d'acord amb aquest ordre: 1 a la diferència menor, 2 a la següent, etc. Si dos problemes de prova coincideixen en el valor de la diferència s'atorga a ambdós el valor mig dels rangs (1.5 si es tractés de la menor diferència observada, per exemple), i així successivament si coincideixen tres o més resultats. A partir d'aquí es defineixen els valors de les sumes dels rangs per cada signe, repartint entre els dos aquells rangs que coincideixen amb diferència nul·la, si és el cas:

$$R^+ = \sum_{d_j>0} rang(D_j) + \frac{1}{2}\sum_{D_j=0} rang(D_j) \tag{6.14}$$

$$R^- = \sum_{d_j<0} rang(D_j) + \frac{1}{2}\sum_{D_j=0} rang(D_j)$$

El valor menor entre R^+ i R^- és conegut com l'estadístic T de Wilcoxon. Si no hi hagués diferència entre ambdós algorismes, els dos valors de R^+ i R^- serien iguals, i coincidirien amb

$$T_0 = \frac{N(N+1)}{4} \tag{6.15}$$

que és el valor esperat per l'estadístic de Wilcoxon. En la comparació sobre la taula de valors crítics corresponents (vegi's, per exemple, [19]) es dóna bàsicament una mesura de la probabilitat que la diferència entre el valor obtingut T i T_0 sigui fruit de l'atzar, i es compara amb el valor de significança α determinat. De la mateixa manera que en el cas del t-test, també es

pot obtenir un valor de p que mesuri la probabilitat que la diferència obtinguda no provingui d'una diferència entre els algorismes, i que permet la seva comparació amb el valor d'α.

La robustesa d'aquesta metodologia ve donada pel fet que les condicions que preveu són molt més febles que en el t-test: no assumeix una distribució normal de les dades, i els *outliers* tenen un menor efecte sobre els resultats, doncs la informació que el test utilitza és més qualitativa que quantitativa. A canvi, quan les condicions que preveu el t-test es compleixen, el test de Wilcoxon és més feble en l'intent de detectar les diferències existents entre els algorismes: veurem en la discussió sobre potència i replicabilitat, al capítol 8 com té menor capacitat de rebutjar H_0 en cas que sigui certa H_1.

Es podria dir que té menys potència però és més segur: en aquesta línia, sovint es diu que el resultat del Wilcoxon és més conservador que el del t-test. El debat sobre quin cal utilitzar, doncs, vindrà del tot determinat pel compliment o no de les condicions prèvies que el t-test demana ([108]), doncs en igualtat de condicions no hi ha dubte: si el t-test es pot aplicar, sempre és la primera opció.

Per tal d'estudiar el domini d'aplicació d'aquest mètode respecte el del t-test, i la potència dels resultats que aporta, recuperem l'exemple que s'ha desenvolupat al final de l'apartat 6.2, en què s'estudien els resultats dels algorismes OAN_08, $OAN\_05\_MAX\_3\_NORM$ i OAN_05_NORM (amb resultats determinats per les magnituds X_1, X_2 i X_3) sobre els problemes de prova corresponents a dominis mèdics utilitzats a l'article [2]. Si es completen els càlculs ja realitzats amb els del test de Wilcoxon, s'obtenen els resultats que es mostren a la taula 6.8.

Aquestes tres comparacions mostren situacions ben diferents: en primer lloc, s'analitza la comparació entre OAN_08 i $OAN\_05\_MAX\_3\_NORM$ (X_1 i X_2). En aquest cas, es complien les suposicions sobre homogeneitat de variàncies però no sobre el caràcter normal de les dades, i per tant no seria possible aplicar el t-test de manera fiable. Tot i així, es pot veure com la conclusió que s'obté pel test de Wilcoxon és la mateixa, i permet rebutjar la hipòtesi nul·la, de la mateixa manera que ho indicava el t-test.

Passa el mateix pel cas de la comparació de $OAN\_05\_MAX\_3\_NORM$ (X_2) i OAN_05_NORM (X_3), en què malgrat no complir-se les condicions que permetein aplicar el t-test, amb el test de Wilcoxon s'obté la mateixa conclusió, i no es pot rebutjar la hipòtesi nul·la.

En canvi, en la comparació entre OAN_08 (X_1) i OAN_05_NORM (X_3), el test de Wilcoxon contradiu la conclusió del t-test, que era errònea

	X_1 **vs** X_2	X_1 **vs** X_3	X_2 **vs** X_3
p-value (t-test)	0.01	0.12	0.31
Rebuig H_0 ($\alpha = 0.05$)	Si	No	No
p-value (S-W)	<0.01	<0.01	<0.01
Normalitat	No	No	No
N	18	18	18
t-test (hom. var.)	0.31	2.62	2.92
$t_{crit}(.05)$	2.37	2.37	2.37
$t_{crit}(.01)$	3.5	3,5	3,5
Rebuig hom. var. ($\alpha = 0.05$)	No	Si	Si
ω^2	<0.01	0.01	<0.01
p-value (Wilcoxon)	<0.01	0.04	0.55
Rebuig H_0 ($\alpha = 0.05$)	Si	Si	No

Taula 6.8: Resum dels resultats per als algorismes comparats, incloent el test de signes de Wilcoxon. El resultat X_1 correspon al de l'agorisme OAN_08, X_2 al de $OAN\_05\_MAX\_3\_NORM$, i X_3 al de OAN_05_NORM. Es veu com en el cas de la comparació $X_1 vs X_3$, el test de Wilcoxon contradiu el resultat obtingut pel t-test, que no era fiable doncs no es complien les condicions que garanteixen el seu domini d'ús.

perquè no es complia la condició de normalitat dels resultats i, per tant, no es podia aplicar amb uns resultats fiables. El test de Wilcoxon permet rebutjar la hipòtesi nul·la, cosa que no passava amb el t-test.

De fet, aquest exemple posa de manifest que els errors més comuns es poden produir en aquells casos propers al nivell de significança escollit α: difícilment el no compliment de les condicions pot posar en dubte valors de p del t-test molt llunyans a α, però en canvi l'efecte del no compliment de les condicions del t-test és molt important quan els valors de p obtinguts s'apropen a α.

Aquest exemple ens permet contextualitzar l'ús del protocol mostrat a la figura 6.2, per a la realització del t-test. Estrictament parlant és correcte, però per valors de p obtinguts amb el t-test que compleixin $p << \alpha$, estiguin molt allunyats d'α la violació d'alguna de les condicions no arriba a tenir efectes determinants sobre la conclusió final: per tant, es conclou recomanant l'aplicació del test de Wilcoxon en aquells casos en que les condicions mostrades en el protocol de la figura 6.2 no es compleixi i, a més, s'obtingui un valor de p d'ordre de magnitud que no sigui molt menor que α.

Les anteriors referències a un relaxament en el domini d'ús del t-test es referien a aquesta conclusió, que a més posen en dubte les opinions d'aquells

que recomanen l'aplicació permanent del test de Wilcoxon, emparant-se en la facilita amb que es viola alguna de les condicions d'ús del t-test. D'acord amb això, el protocol per a l'aplicació del t-test pren la forma definitiva que es mostra a la figura 6.3, on s'inclou la informació sobre el valor obtingut de p.

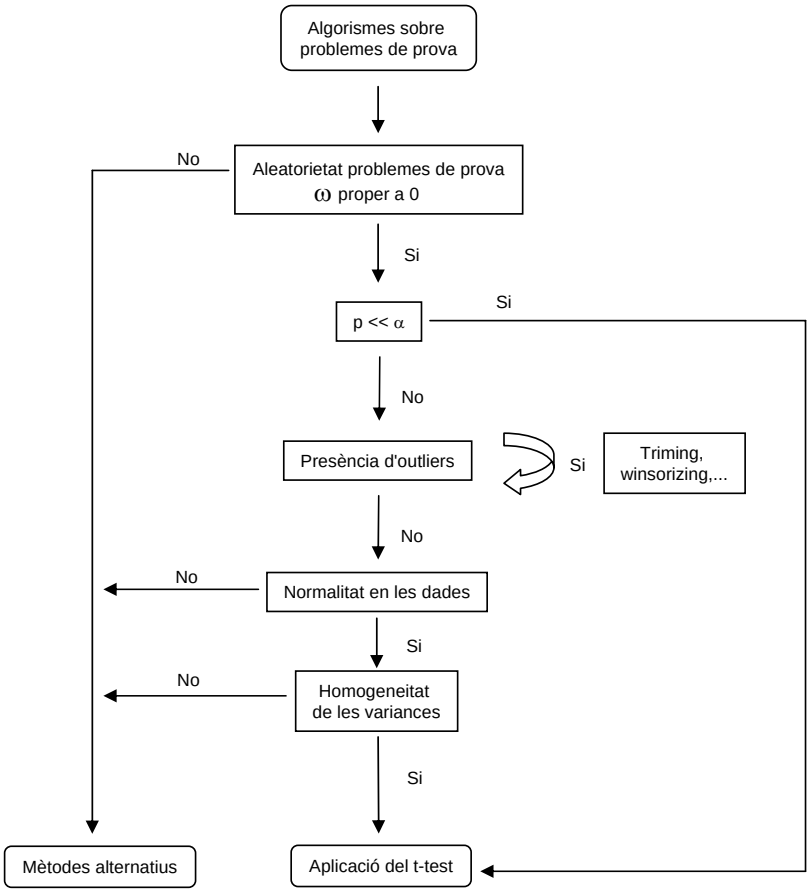

Figura 6.3: Proposta de protocol per a la correcta aplicació del t-test en la comparació de resultats de dos algorismes sobre un conjunt d'N problemes de prova.

6.3.3 Test de signe binomial i de McNemar

Una altra aproximació no-paramètrica al problema és el conegut com a test de signe binomial, que fa un pas més enllà en el camí iniciat pel test de Wilcoxon: si aquell considerava tot just informació qualitativa sobre la diferència entre

els resultats obtinguts pels dos algorismes, aquest redueix aquesta informació a la mínima possible. Tant és així, que esdevé un simple contatge de guanys, empats i pèrdues entre els dos algorismes comparats, sense preocupar-se per la magnitud del guany o de la pèrdua.

Si la hipòtesi nul·la és certa, els algorismes comparats tenen comportament equivalents i, per tant, és d'esperar que cadascun d'ells guanyi aproximadament en la meitat dels problemes de prova. O, dit d'una altra manera, ambdós tenen la mateixa probabilitat de comportar-se millor que l'altre en cada problema, i per tant la variable que mesura el número de guanys segueix una distribució binomial equiprobable ([43]), que a més es pot aproximar per una distribució normal de mitjana $N/2$ i dispersió $\sqrt{N}/2$ per un número de problemes de prova elevat ([95]). D'acord amb això, la probabilitat que un algorisme sigui millor que l'altre un número mínim de M vegades sobre N problemes de prova és igual a

$$P(guanys \geq M) = \sum_{j=M}^{N} \left(\begin{array}{c} N \\ j \end{array} \right) \frac{1}{2}^{j} \frac{1}{2}^{N-j} \tag{6.16}$$

El test binomial és encara més conservador que el de Wilcoxon, en el sentit que utilitza menys informació i això el converteix en més robust (només un conjunt de problemes de prova no aleatòriament escollits podria invalidar-ne la conclusió), però a la vegada és menys capaç en detectar les diferències significatives entre els algorismes, en cas que aquestes existeixin.

És important interpretar correctament la manera d'utilitzar aquest test, i no caure en un error relativament habitual: alguns autors creuen que les conclusions seran més fiables si només es tenen en compte els guanys "significatius" entre els algorismes, és a dir, aquelles diferències en el comportament que són estadísticament significants (per exemple, aplicant un t-test sobre les dades obtingudes en el procés de *cross-validation* seguit per avaluar l'error de test). En la referència anterior de les edicions 1999-2003 de l'*International Conference on Machine Learning*, per exemple, aquest és un error comès aproximadament per un 10% dels articles analitzats.

Es pot entendre fàcilment el perquè de l'error portant-ho al límit: suposem N problemes de prova, sobre els quals l'algorisme A_1 sempre fos millor que l'algorisme A_2, tot i que les diferències mai fessin possible que un t-test sobre els *folders* del *cross-validation* detectés una diferència significativa. Sembla bastant evident que, per valors grans de N, amb tota seguretat es podria rebutjar la hipòtesi nul·la H_0, i de fet així ho farien tots els test de comparacions simples que apliquéssim. Ara bé, un contatge de guanys i pèr-

dues "significatives" permetria mantenir la suposició que ambdós algorismes es comporten de manera similar, doncs no se'n trobaria cap.

Com alternativa al test binomial, el test de McNemar és útil quan els resultats es mostren en categories: és a dir, quan els resultats no són numèrics o ordinals, sinó nominals. Habitualment, això es produeix en problemes que no són de l'àmbit d'estudi del present treball, i per aquest motiu no es tractarà en aquest punt. Es poden trobar referències útils a [109] o [110].

6.3.4 Matriu de guanys

La simplicitat del test binomial fa que en algunes ocasions s'extrapoli la seva utilització per problemes de comparacions múltiples, en els quals es compara el comportament de M algorismes assajats sobre N problemes de prova, amb $M > 2$ (cas que s'estudiarà, àmpliament, al capítol següent).

L'extrapoliacó consisteix en separar el problema de la comparació de M algorismes en tots els $M(M-1)/2$ problemes de comparació simples per parelles possibles, entre dos algorismes. A partir d'aquí, es defineix una matriu G de guanys, on l'element G_{ij} és el número de problemes de prova en què l'algorismes i obté un millor resultat que l'algorisme j. A partir d'aquesta definició es veu com els termes G_{ii} no tenen sentit (habitualment es troben matrius amb les diagonals buides), i com sempre es compleix $G_{ij} \leq N$.

El problema d'aquest plantejament és que permet treure menys conclusions que les que habitualment se n'obtenen. Per exemple, si definim G_i com els guanys de l'algorisme i, segons l'expressió

$$G_i = \sum_{j=1}^{M} G_{ij} \tag{6.17}$$

sovint es treballa amb aquest valor com una magnitud que expressa la bondat de l'algorisme i respecte els altres. Aquesta interpretació pot ser certa, però no disposem d'eines estadístiques que permetin discutir si els valors obtinguts per G_i determinen diferències significatives, a menys que es redueixi el problema a una comparació simple de l'algorisme i respecte els altres. Per això, les conclusions sovint són poc estrictes, i acaben amb frases del tipus "l'algorisme A_1 guanya en G_{12} ocasions a l'algorisme A_2, que és millor que A_3 un total de G_{23} vegades, i només pitjor en G_{32} ocasions".

També en aquest cas és comú l'error de construir una matriu de guanys només amb aquells casos en què la millora és significativa, d'acord amb el

comentat a l'apartat immediatament anterior. Per demostrar la magnitud de l'error en aquesta interpretació, s'han analitzat els resultats publicats a [4]. En aquest article, els autors comparen el comportament de $M = 8$ algorismes diferents sobre $N = 15$ problemes de prova que provenen del repositori UCI ([3]) i de dades pròpies ([51], [111] i [112]). Els resultats es poden veure a la taula 6.9.

Prob.	**ADI**	**ADI1**	**ADI2**	**ADI3**	**ADI4**	**ADI5**	**C4.5**	**IB1**
bpa	63.7	63.7	63.9	62.6	63.3	63.2	68.4	64.5
bps	80.6	80.7	80.7	79.6	80.6	80.0	80.1	83.2
bre	95.6	95.8	96.0	95.7	95.8	95.9	95.4	96.0
gls	66.4	66.5	67.9	67.9	67.8	66.5	65.8	66.3
h-s	80.4	80.2	80.7	79.8	80.5	80.6	76.3	74.1
ion	91.6	90.9	92.2	91.7	92.7	92.0	89.8	86.9
lrn	68.1	68	68.6	67.8	68.9	69.0	68.6	61.4
mmg	65.0	66.1	66.0	67.8	67.0	67.8	64.8	63.5
pim	74.4	75.1	74.7	74.3	75.3	74.4	73.1	70.3
son	74.6	73.2	73.1	72.3	74.3	73.5	71.5	87.3
thy	91.9	92.0	91.4	91.6	92	91.5	92.6	96.8
veh	66.0	66.4	66.1	65.6	66.7	66.5	73.6	69.4
wdbc	93.8	93.7	93.8	93.9	94.0	93.7	93.7	95.6
wine	92.7	92.5	92.2	92.6	93.0	92.2	94.1	95.6
wpbc	75.7	75.5	76.1	75.9	77.1	76.8	73.7	68.8
Mitjana	78.7	78.7	78.9	78.6	79.3	78.9	78.8	78.6

Taula 6.9: Resultats de [4], amb l'aplicació de les cinc variants de l'algorisme ADI assajades (veure la informació addicional que s'exposa a l'apartat **??**, el propi algorisme ADI i els algorismes $C4.5$ i $IB1$.

Sobre aquests resultats s'hi van aplicar tot un seguit de t-test amb un nivell de significança $\alpha = 0.01$, obtenint la matriu de guanys i pèrdues "significants" que es pot observar a la taula 6.10. En canvi, a la taula 6.11 es pot observar la matriu de guanys tal i com estan definits realment en el test binomial. S'hi poden observar diferències importants, és a dir, errors que una taula de guanys "significants" pot provocar.

Per començar, només hi ha dues comparacions en què el resultat és significatiu, d'acord amb el test de signes binomial: tenint en compte que cada parella d'algorismes són comparats sobre els 15 problemes de prova proposats, ha d'haver-hi un mínim de 12 guanys per a què es pugui considerar

Algorisme	**ADI**	**ADI1**	**ADI2**	**ADI3**	**ADI4**	**ADI5**	**C4.5**	**IB1**	**Total**
Original ADI		0	0	0	0	0	1	3	4
New ADI1	0		0	0	0	0	0	4	4
New ADI2	0	1		1	0	0	1	4	7
New ADI3	1	0	1		0	0	1	3	6
New ADI4	2	2	0	3		1	1	4	13
New ADI5	1	0	0	2	0		1	4	8
C4.5	2	1	1	2	2	2		0	10
IB1	3	4	4	3	3	3	2		22
Total	9	8	6	11	5	6	7	22	

Taula 6.10: Resum dels resultats de [4] amb l'aplicació d'un t-test amb significança estadística $\alpha = 0.01$. Cada valor indica quantes vegades l'algorisme de la fila obté un millor resultat "significatiu" que l'algorisme de la columna.

Algorisme	**ADI**	**ADI1**	**ADI2**	**ADI3**	**ADI4**	**ADI5**	**C4.5**	**IB1**	**Total**
Original ADI	-	7	3	9	2	6	10	7	44
New ADI1	7	-	6	9	2	5	9	7	45
New ADI2	11	8	-	10	5	8	10	7	59
New ADI3	6	6	4	-	2	4	9	7	38
New ADI4	12	11	10	13	-	11	11	7	75
New ADI5	8	8	6	10	4	-	9	7	52
C4.5	5	5	4	6	4	5	-	8	37
IB1	8	8	7	8	8	8	7	-	54
Total	57	53	40	65	27	47	65	50	

Taula 6.11: Matriu de guanys dels resultats de [4]. Cada valor indica quantes vegades l'algorisme de la fila obté un millor resultat que l'algorisme de la columna.

una diferència significativa amb $\alpha = 0.05$, d'acord amb el definit a l'equació 6.16. És a dir, tan sols l'algorisme $ADI4$ obté resultats significativament millors, en concret en la comparació amb els ADI i $ADI3$. Les altres diferències observades no són estadísticament significatives al nivell α.[3]

En aquesta línia, si es vol obtenir una certa idea del conjunt sense aplicar un test de comparació múltiple, es veu clarament com és precisament l'algo-

[3]La publicació original, com s'ha dit, utilitzava un nivell de significança $\alpha = 0.01$. D'acord amb el que s'ha comentat ja a l'apartat 5.1, en relació als error "Tipus II", és excessiu rebaixar tant el valor d'α, si no es comprova el control sobre aquest error. Per això tots els càlculs que es realitzen en aquest treball sobre aquelles dades es fan amb $\alpha = 0.05$.

risme $ADI4$ el millor de tots, amb 72 guanys per només 27 casos en què es comporta pitjor. En canvi, l'algorisme $IB1$ deixa de mostrar segons aquests resultats els excel·lents resultats que mostrava en la matriu de guanys "significatius". O un darrer exemple més: malgrat el $ADI4$ només guanya 2, 2, 3 i 1 vegada "significativament" als algorismes ADI, $ADI1$, $ADI3$ i $ADI5$, respectivament, un test de Wilcoxon ens permet rebutjar en tots els casos la hipòtesi H_0, i per tant concloure que, en la comparació per parelles, l'algorismes $ADI4$ és significativament millor que aquests altres quatre, conclusió a la qual en cap cas ens permetria arribar la taula 6.10.

Més enllà d'aquestes primeres afirmacions particulars, però, hi ha dues consideracions a fer: d'una banda, que considerar els guanys "significatius" pot provocar conclusions equivocades o parcials; d'altra banda, que quan es té un conjunt d'algorismes major a 2, sembla clar que calen metodologies més complexes que l'extrapolació de les comparacions simples, que en el fons és el que es mostra a les matrius de guanys, i que poden induir a conclusions molt parcials. Aquesta mancança es resoldrà al capítol següent, i en l'apartat **??** es posarà de relleu l'error de les conclusions a què es poden arribar amb una anàlisi com la que s'ha fet en aquest apartat.

6.4 Resum

Després d'haver posat les bases de l'anàlisi dels resultats, en aquest capítol s'ha iniciat la discussió dels test d'inferència estadística per al cas de comparacions simples, en què es compara el comportament de dos algorismes, a partir dels resultats obtinguts després del seu assaig sobre una col·lecció de problemes de prova.

En primer lloc s'ha fet una breu explicació del t-test, la principal eina paramètrica per a la comparació simple. En aquest cas, l'estudi de l'efecte dels *outliers*, la normalitat dels resultats obtinguts i l'homogeneïtat de les variàncies han estat els elements discutits per arribar a proposar un protocol d'aplicació del t-test, en funció del compliment de totes aquelles restriccions que emmarquen l'obtenció de resultats fiables. La determinació del domini d'ús s'ha mostrat en un cas sobre problemes de prova de domini mèdic, amb algorismes provinents de diverses variacions del SOMCBR. Una primera conclusió ha estat la gran dificultat per al compliment de totes les condicions que determinen aquest domini d'ús, i s'ha obert la porta al relaxament d'alguna d'elles.

A continuació, s'han estudiat les diverses variants no-paramètriques possi-

bles, per aquells problemes en què el t-test no es pugui aplicar amb garanties. D'una banda, s'ha realitzat un estudi complert del test de Wilcoxon, relacionant-ho amb els resultats del t-test sobre els problemes de dominis mèdics abans esmentats. Aquests resultats han permès mostrar un possible relaxament sobre les condicions que determinen el domini d'ús del t-test, conduint al protocol final exposat a la figura 6.3.

D'altra banda, s'ha estudiat també el funcionament del test binomial i la seva aplicació en la construcció d'una matriu de guanys, en un intent de generalitzar a M algorismes el presentat per una comparació simple. A partir de resultats també publicats, s'han mostrat els errors que es poden produir si no s'interpreta correctament aquesta matriu, i s'ha fet patent la gran dificultat d'extrapolar a casos complexes les tècniques de comparació simple.

Capítol 7

Comparació múltiple de resultats

"Althought the difference between $\alpha_{FW} = 0.05$ and $\alpha_{FW} = 0.10$
may seem trivial, a result that is declared significant
is more likely to be submitted and/or accept for publication"
David J. Sheskin, [19]

"Many otherwise excellent and innovative machine learning papers end (...)
as if the tests for multiple comparisons,
such as ANOVA and Friedman test, are yet to be invented."
J. Demsar, [9]

En el capítol anterior s'ha procedit a l'anàlisi de les diferents metodologies que permeten la comparació del comportament de dos algorismes, incloent-hi l'establiment d'un protocol per a la seva aplicació, insinuant-hi la relació d'aquest amb magnituds com la replicabilitat o la potència d'un test, que serà mostrada al capítol següent. El que allí s'ha exposat, juntament amb els capítols precedents, resol el problema de la comparació simple de resultats.

En aquest capítol es repetirà un plantejament similar, però ara pel cas de la comparació múltiple, és a dir, quan el número d'algorismes a comparar és major que 2. Es presentarà el test paramètric de referència, amb un exhaustiu estudi sobre el seu domini d'ús, i els tests per aplicar posteriorment i discutir l'existència de diferències significatives, en aquells casos en que s'hagi rebutjat la hipòtesi nul·la. Això es farà tenint en compte l'anàlisi del nivell de significança determinat globalment, que admet diverses modificacions quan es passa a la comparació entre un número d'algorismes menor que els inicials.

Posteriorment, s'analitzaran les alternatives no paramètriques (en cas que no es compleixin les condicions que permetin l'ús d'un test paramètric), tant per a analitzar la hipòtesi nul·la global com per a determinar quina o quines són les diferències significatives entre els algorismes. Totes aquestes consideracions es reuneixen en un protocol d'actuació amb que es conclou el capítol.

7.1 Plantejament

Quan el número d'algorismes que es comparen és superior a 2, el primer que cal determinar és quina és la hipòtesi nul·la sobre la qual es discutirà: habitualment, el primer plantejament és intentar rebutjar la condició d'igualtat de comportament de tots els algorismes, per centrar-se posteriorment en quins són aquells significativament diferents. Aquest plantejament, i els efectes que comporta sobre la determinació del nivell α de significança estadística, és el que es discuteix a l'apartat 7.2.

A continuació, en l'apartat 7.3 es discuteixen les tècniques d'inferència vàlides per enfocaments paramètrics. S'exposa l'anàlisi de variàncies explicant com es realitzen els càlculs de l'estadístic corresponent i mostrant els resultats sobre un exemple, i es discuteix posteriorment el domini d'ús d'aquesta metodologia (apartat 7.3.2).

Entenent que la principal alternativa serà recórrer a un test no paramètric, s'estudien amb molta profunditat el compliment de les condicions que es requereixen per a l'ús del test paramètric, proposant un protocol que faciliti el càlcul d'aquestes condicions i l'anàlisi de com de crític n'és l'incompliment de cada una d'elles.

L'apartat sobre els test paramètric acaba amb l'exposició dels test que, posteriorment a l'hipotètic rebuig d'H_0, poden permetre determinar quina o quines comparacions simples són responsables d'aquest rebuig i, per tant, mostren quins algorismes tenen un comportament diferent. L'anàlisi es centra en determinar com treballar amb les diferents alternatives que existeixen per al càlcul del que es definirà com a distància crítica.

A continuació (apartat 7.4) s'estudien les alternatives no paramètriques, començant per exposar com calcular l'estadístic que aporta informació sobre el possible rebuig de la hipòtesi nul·la global. D'igual manera que amb els test paramètrics, s'estudien a continuació les possibilitats existents per als test a posteriori, determinant-ne la utilitat en cada cas i mostrant una representació gràfica útil per a l'expressió de les conclusions, en la línea del proposat a l'apartat 5.2.

Finalment, el capítol acaba (apartat 7.5) amb un protocol per a l'aplicació dels diferents test vàlids per a problemes de comparació múltiple que inclou els estudis fets sobre el domini d'ús dels test paramètrics, la utilització correcta del concepte de distància crítica i les diferents opcions de test a posteriori, determinant sempre una resposta sobre les hipòtesis plantejades i fent un plantejament de tall conservador.

7.2 Hipòtesis i control de la precisió

En el darrer apartat del capítol anterior, s'ha vist una de les múltiples formes d'extrapolació de test de comparació simple per a l'anàlisi múltiple de resultats: les matrius de guanys i pèrdues representades, amb els sumatoris de files i columnes corresponents, donen informació sobre el comportament dels M algorismes assajats sobre els N problemes de prova, però limiten unes conclusions clares i generals. Malgrat això, aquestes matrius són molt sovint utilitzades en els treballs desenvolupats en el nostre àmbit de coneixement.

7.2.1 Plantejament de les hipòtesis

Partint de les conclusions amb què s'ha acabat el capítol anterior, cal començar posant de manifest que no és vàlida una generalització de les comparacions simples. De fet, existeix una metodologia utilitzada sovint que es basa en això: suposant M algorismes sobre N problemes de prova, es realitzen tots les comparacions simples (per parelles) possibles, un número igual a

$$(M-1)+(M-2)+...+1=\frac{M(M-1)}{2} \tag{7.1}$$

Un cop fetes aquestes comparacions amb un nivell de significança estadística α, el problema acaba amb conclusions del tipus "l'algorisme A_1 és significativament millor que els A_4 i A_6, mentre que els A_2 i A_3 són significativament millors que el A_5". Aquestes conclusions tenen molt poc sentit, com veurem posteriorment no són del tot fiables i, sobretot, no donen una resposta complerta a la pregunta de si hi ha algun algorisme millor que la resta, després d'avaluar-los sobre el conjunt de problemes de prova de què es disposa.

De fet, la clau està precisament en quina pregunta és la correcta o, dit de manera més adequada, quina hipòtesi nul·la es vol sotmetre a test. Els test

de comparació múltiple es fan sobre una hipòtesi H_0, i la seva alternativa H_1, comuna per a tots els algorismes, i plantejada com segueix:

H_0 : Tots els algorismes tenen el mateix comportament, i les diferències observades són fruit de l'atzar

H_1 : Dins el conjunt dels M algorismes, hi ha com a mínim dos algorismes que presenten comportaments diferents

L'hipotètic rebuig d'aquesta hipòtesi nul·la H_0 no aportaria informació sobre quins són els algorismes que provoquen aquestes diferències, però com a mínim permetria iniciar un estudi més profund sobre la base de l'existència d'almenys una diferència significativa. De fet, fins i tot podria ser que cap algorisme mostrés un comportament significativament millor que cap altre, sinó que la diferència en el conjunt vingués provocada per una comparació entre més d'un algorisme (per exemple, que A_1 fos millor que la mitjana de A_2 i A_3). De fet, el rebuig de la hipòtesi nul·la que es plantejarà afirma que existeix com a mínim una diferència significativa d'entre totes les possibles comparacions que es podrien arribar a fer, ja siguin per parelles o més complexes (com es mostra a [106], [113] o [114]).

7.2.2 Control del nivell de significança

Un cop establert el context en què es treballarà, cal prestar atenció a un dels concepte més importants en tot aquest capítol: la significança estadística múltiple, habitualment α_{FW}, per l'anglès *family-wise*. Fins ara, amb les comparacions simples, establíem un nivell de significança (que a partir d'ara escriurem com α_{PC}, per l'anglès *pairwide comparison*) que determinava l'error "Tipus I" que s'aspirava a cometre (és a dir, el límit que es posava a la probabilitat de l'error comès al rebutjar la hipòtesi H_0, en el cas que ambdós algorismes comparats tinguessin realment el mateix comportament). Ara bé, si la comparació que es realitza és múltiple, aquest concepte ha de patir algunes variacions per continuar essent vàlid.

Tal i com està definida ara H_0, α_{FW} serà la probabilitat que almenys una de les comparacions possibles dugui a una diferència significativa. Si imaginem que es duen a terme c comparacions simples (les $M(M-1)/2$ entre parelles, per exemple, o d'altres més complexes), α_{FW} es pot calcular com el complementari de la probabilitat de no tenir cap error de "Tipus I":

$$\alpha_{FW} = 1 - P[\text{cap error "Tipus I"}] = 1 - (1 - \alpha_{PC})^c \qquad (7.2)$$

Aquesta expressió és exacta en el cas que totes les comparacions siguin independents entre elles i depèn només del valor que prengui α_{PC} i del número de comparacions simples que es facin. Sigui com sigui, l'important és veure que sempre es complirà la relació

$$\alpha_{FW} > \alpha_{PC} \tag{7.3}$$

i, per tant, que garantir un cert valor de significança per cada comparació per parelles, per exemple, no suposa realitzar una anàlisi pel mateix nivell de significança per al conjunt.

D'acord amb això, el lògic seria efectuar el test múltiple sobre la hipòtesi nul·la global amb un habitual $\alpha_{FW} = 0.05$ i, en cas que es detecti alguna possible diferència significativa, a continuació calcular el corresponent α_{PC} per dur a terme les comparacions simples i esbrinar quin o quins algorismes provoquen aquesta diferència detectada. Aquest plantejament, però, no és tan simple d'aplicar, doncs per valor de c elevats el valor de α_{PC} tendeix ràpidament a 0 i, aleshores, l'error de "Tipus II" (la probabilitat de no rebutjar la hipòtesi H_0 quan realment hi ha una diferència significativa) creix massa com per considerar les conclusions fiables.

Cal trobar, per tant, un punt mig en aquest equilibri i determinar una bona estratègia per al que sovint s'anomena el control de l'α_{FW} que, tot sigui dit, pràcticament mai és dut a terme en els treballs que es publiquen habitualment al nostre àmbit de coneixement.

Malgrat ser aquesta una qüestió ben treballada des d'un punt de vista teòric, no hi ha pas un clar consens entre la nostra comunitat sobre la manera òptima d'actuar en el control de l'α_{FW} ([19]). Diversos exemples i contraexemples es succeeixen en la bibliografia ([113], [114]) intentant mostrar les virtuts de les diferents estratègies possibles. Vistes les propostes, sembla que hi ha un concepte clau que centra la discussió: la diferència entre una comparació planificada a priori o una decidida a posteriori (*planned* o *unplanned*).

Una comparació planificada succeeix quan el propi disseny de l'experiment es realitza pensant en comparar la bondat relativa de dos algorismes (X_1 vs X_2, una comparació simple com les realitzades fins ara) o de dos grups d'algorismes: l'anomenarem comparació complexa, que no múltiple[1] (X_1 vs $(X_2+X_3)/2$, per exemple). En cap d'aquests casos cal ajustar el valor de α_{FW}

[1] Convé recordar un cop més que ens referim a una comparació complexa quan es tracta de comparar dues mesures (un tipus de comparació simple, com també ho són les comparacions per parelles), mentre que una comparació múltiple són aquelles que n'involucren més de dues.

i, de fet, es treballa amb α_{FW} com en tot el capítol anterior, amb l'habitual valor de 0.05.

En canvi, una comparació no planificada o decidida a posteriori és aquella que es realitza després de fer la comparació múltiple sobre el conjunt dels M algorismes, amb un determinat valor de α_{FW}, i havent-hi trobat una diferència significativa: com a mínim una de les possibles comparacions simples mostrarà una diferència significativa. En aquest cas, a posteriori es poden efectuar tot un conjunt de comparacions, i això donarà lloc als coneguts com a test *post-hoc*. És en aquest punt en què no hi ha un acord entre la comunitat: des d'aquells que consideren inexcusable el control de α_{FW} per reduir el α_{PC} utilitzat en aquest test *post-hoc*, fins a d'altres que ho consideren només possible.

La nostra tesi, que estarà present en tot el text, és que l'important no és tant reduir o no el valor de α_{PC} segons la fórmula 7.2, sinó conèixer plenament què implica la utilització d'una o altra estratègia, de les quals s'obtindran resultats diferents. Per aquest motiu, aquest capítol es divideix en dues parts ben diferenciades. D'una banda, uns primers apartats en què s'exposaran els test per executar comparacions múltiples (paramètrics i no paramètrics), i d'altra banda uns apartats on s'exposaran les diferents propostes de test *post-hoc* (tant en el cas del paramètric com en el no-paramètric), amb una anàlisi detallada del domini d'aplicació de cada resultat en funció del tractament realitzat sobre α_{FW} i α_{PC}. El protocol proposat amb què finalitzarà el capítol inclourà totes aquestes consideracions.

7.3 Test paramètrics

L'estudi de les diferents opcions en quant al test d'inferència estadística per a les comparacions múltiples es farà seguint un esquema similar al que ja s'ha fet per a les comparacions simples. El primer nivell d'anàlisi el dona el fet que les dades que s'obtinguin compleixin un conjunt de restriccions i permetin l'aplicació d'un test paramètric. En cas contrari, cal procedir a l'aplicació d'una alternativa no paramètrica, que s'estudiarà a l'apartat 7.4.

A continuació s'estudiarà el domini d'ús de l'anàlisi de variàncies, la principal metodologia de test paramètric per comparacions múltiples, i aquells test que es poden aplicar a posteriori, en cas que el valor de l'estadístic calculat permeti rebutjar la hipòtesi nul·la, H_0.

7.3.1 Anàlisi de variàncies (ANOVA)

El test paramètric més habitual quan es tracta de fer comparacions múltiples és l'anàlisi de variàncies, també conegut com ANOVA ([115]). Aquest mètode es basa en una comparació entre les diferents fonts que poden provocar la variabilitat observada a les dades, una variació que pot venir provocada per la diferència real de comportament entre els algorismes, per factors propis de l'experimentació o per d'altres que no es poden atribuir als propis algorismes.

Tal i com s'ha dit abans, la hipòtesi nul·la afirma que no hi ha diferència entre el conjunt dels algorismes, a partir de les dades obtingudes de l'aplicació sobre la col·lecció de problemes de prova de què es disposa. En cas de rebutjar-se caldrà treballar amb test a posteriori per determinar quin o quins algorismes són els més ben comportats. La relació entre els factors que poden introduir variació en els resultats determinarà el rebuig o acceptació d'aquesta hipòtesi nul·la global.

La variabilitat total a les dades obtingudes es pot atribuir a tota una sèrie de factors: la deguda pròpiament a les diferències de comportament entre els algorismes (BA, de *between algorithms*); la deguda a les diferències entre els problemes de prova (BD, de *between datasets*) i aquella que ve determinada per les condicions experimentals o l'atzar (*res*, per *residual*). Si, d'acord amb les nomenclatures utilitzades fins al moment, definim $X_{i,j}$ com la mesura obtinguda d'aplicar l'algorisme i sobre el problema de prova j, es pot calcular la variabilitat total (SS_T) de les dades obtingudes per la següents expressió:

$$SS_T = SS_{BA} + SS_{BD} + SS_{res} = \sum_{i=1}^{M}\sum_{j=1}^{N} X_{i,j}^2 - \frac{1}{NM}\left(\sum_{i=1}^{M}\sum_{j=1}^{N} X_{i,j}\right)^2 \quad (7.4)$$

De manera similar, es poden definir les variabilitats entre els algorismes (SS_{BA} i entre els problemes de prova (SS_{BD}):

$$SS_{BA} = \sum_{i=1}^{M} \frac{\left(\sum_{j=1}^{N} X_{i,j}\right)^2}{N} - \frac{1}{NM}\left(\sum_{i=1}^{M}\sum_{j=1}^{N} X_{i,j}\right)^2 \quad (7.5)$$

$$SS_{BD} = \sum_{j=1}^{N} \frac{\left(\sum_{i=1}^{M} X_{i,j}\right)^2}{M} - \frac{1}{NM}\left(\sum_{i=1}^{M}\sum_{j=1}^{N} X_{i,j}\right)^2$$

i d'aquí s'obté la mesura per la variabilitat residual (SS_{res}):

$$SS_{res} = SS_T - SS_{BA} - SS_{BD} \quad (7.6)$$

Totes aquestes mesures es normalitzen amb els graus de llibertat (df) corresponents per a cada cas:

$$df_{BA} = M - 1 \tag{7.7}$$

$$df_{BD} = N - 1$$

$$df_{res} = (M-1)(N-1)$$

de tal manera que determinen les mesures per a la variabilitat normalitzades:

$$MS_{BA} = \frac{SS_{BA}}{df_{BA}} \tag{7.8}$$

$$MS_{BD} = \frac{SS_{BD}}{df_{BD}}$$

$$MS_{res} = \frac{SS_{res}}{df_{res}}$$

L'estadístic F s'obté de dividir la mesura normalitzada de variabilitat que correspon a la diferència entre els algorismes (MS_{BA}), i la corresponent als efectes propis de l'experimentació i l'atzar (MS_{res}):

$$F = \frac{MS_{BA}}{MS_{res}} \tag{7.9}$$

Si MS_{BA} és molt més elevat que MS_{res}, es considerarà altament probable rebutjar la hipòtesi nul·la, doncs la variabilitat observada serà culpa en bona part de les diferències de comportament entre els algorismes. En canvi, si ambdós valors són propers, el quocient serà proper a 1 i no es rebutjarà H_0, en tant que la contribució de la diferència entre els algorismes serà similar a la produïda per les pròpies condicions experimentals del problema, o directament per elements aleatoris presents a l'algorisme o al procediment per obtenir el valor $X_{i,j}$. L'estadístic F segueix la que es coneix precisament com a distribució F i està perfectament tabulada en la majoria dels textos de referència en estadística matemàtica.

Un bon exemple es pot trobar en els resultats publicats a [5]. En aquest treball, entre d'altres coses, s'analitzen els efectes del número de veïns propers (*k nearest-neighbours*) que s'utilitzen a la fase de recuperació d'un esquema CBR (*Case Based Reasoning*), prèvia aplicació d'un procés de clusterització de la memòria de casos sense supervisió (SOMCBR, definit àmpliament a [21]). Els resultats de la taula 7.1 mostren el percentatge d'error en la classificació per un conjunt de 13 problemes de prova de domini mèdic i de dues

classes. Els algorismes assagen diferents valors de k per una estratègia clàssica de CBR, i per dues estratègies SOMCBR amb diferències al procediment de recuperació: recuperant en funció de majories (*vot*) o d'una certa funció de pertinença (*per*).

Dataset	CBR			**SOMCBR-vot**			**SOMCBR-per**		
	$k=1$	$k=3$	$k=5$	$k=1$	$k=3$	$k=5$	$k=1$	$k=3$	$k=5$
Biopsia	16.85	17.33	15.97	23.79	21.58	21.50	24.01	10.23	7.14
mamografies	37.50	35.19	33.33	37.22	35.82	36.51	38.60	23.80	18.51
ddsm2c1	16.97	15.17	15.57	17.66	15.72	14.40	18.43	9.71	7.37
ddsm2c2	33.73	34.13	30.74	38.15	37.46	37.12	37.69	25.92	22.60
ddsm2c3	32.14	30.94	32.73	36.78	36.56	36.24	36.66	26.31	17.34
ddsm2c4	22.16	19.96	19.36	25.70	22.27	21.52	25.55	14.53	17.34
mias3c2c1	19.88	22.05	22.36	27.39	26.56	28.42	27.30	16.89	10.01
mias3c2c2	12.11	9.63	10.25	15.72	17.09	17.50	16.51	9.04	7.12
mias3c2c3	25.78	27.95	28.26	32.84	33.00	32.31	32.65	24.12	19.45
miasbi2c1	12.19	10.31	12.19	20.23	20.47	21.43	20.12	9.61	8.00
miasbi2c2	24.06	22.50	20.94	30.24	26.12	26.15	31.32	15.66	10.98
miasbi2c3	21.25	16.56	16.88	26.64	24.42	26.15	26.72	14.14	8.05
miasbi2c4	9.69	9.69	10.62	13.92	12.88	13.07	14.44	7.68	6.87
Mitjana	21.87	20.88	20.71	26.64	25.38	25.56	26.92	15.97	12.37

Taula 7.1: Resultats de l'error de classificació (en %) de l'aplicació de les 9 estratègies definides al text sobre els 13 problemes de prova. Tots ells són de domini mèdic i els seus elements pertanyen a dues classes.

En aquesta ocasió, la voluntat és estudiar l'efecte de l'augment del valor de k (els veïns més propers que es tenen en compte per decidir la classe de pertinença en la fase de recuperació del CBR), i veure si aquest efecte és constant per les tres estratègies estudiades i presentades en el citat article. De l'observació dels valors mitjos descrits a la taula 7.1 es pot veure com per l'estratègia $SOMCBR-per$ l'error es redueix en gran magnitud a mesura que augmenta k. En canvi, en les altres dues estratègies, sembla clar que $k=3$ i $k=5$ es comporten millor que $k=1$, però no si es pot afirmar això per un cert grau α de significança estadística.

Un anàlisi de variàncies tal i com s'ha definit serveix per respondre a aquesta qüestió, i aporta els resultats que es mostren a la taula 7.2. El test s'ha realitzat per cada una de les tres estratègies, i la hipòtesi nul·la global (H_0) sosté que els tres valors possibles de k porten a tres algorismes amb igual capacitat de classificació.

	CBR	**SOMCBR-vot**	**SOMCBR-per**
SS_T	2688.76	2447.05	3252.97
SS_{BA}	10.25	11.99	1493.94
SS_{BD}	2632.25	2407.59	1608.82
SS_{res}	46.26	27.48	150.21
df_{BA}	2	2	2
df_{BD}	12	12	12
df_{res}	24	24	24
MS_{BC}	5.12	5.99	746.97
MS_{BS}	219.35	200.63	134.07
MS_{res}	1.93	1.15	6.26
Estadístic F	2.66	5.23	119.35
$F_{crit}(\alpha = 0.05)$	3.40	3.40	3.40
$F_{crit}(\alpha = 0.01)$	5.61	5.61	5.61

Taula 7.2: Anàlisi de variàncies per cada una de les tres estratègies assajades a [5]. Com mostren els valors obtinguts per l'estadístic F, en dos dels tres casos es poden rebutjar les hipòtesi nul·les, H_0, per un nivell de significança $\alpha = 0.05$.

En aquest problema, H_0 no pot ser rebutjada per la primera estratègia (CBR), però sí que ho pot ser clarament per la tercera ($SOMCBR-per$) i també per la segona ($SOMCBR-vot$), tot i que en aquest darrer cas només per un nivell $\alpha_{FW} = 0.05$. D'acord amb allò expressat als apartats anteriors, la conclusió a la qual permet arribar l'ANOVA és que en un plantejament CBR no és pot demostrar que l'augment de k faci variar significativament la precisió del classificador, i que en els altres dos plantejaments l'augment de k afecta aquesta precisió. Faran falta test posteriors per determinar quins algorismes són realment millors que els altres, en els casos en que s'ha rebutjat H_0, i suposant que es compleixin les condicions per les quals es pot aplicar l'ANOVA.

7.3.2 Domini d'ús de l'anàlisi de variàncies

De la mateixa manera que en el cas del t-test (el test paramètric per a comparacions simples), l'anàlisi de variàncies també és un test paramètric, i per tant la seva aplicació té sentit només quan es compleixen tot un conjunt de condicions, establertes sobre les dades del problema i la col·lecció de prob-

lemes de prova que s'utilitza. En aquest cas es parla de tres condicions:

1. Selecció aleatòria: la mostra de N problemes de prova ha estat seleccionada a l'atzar d'entre la població de problemes de prova existents.

2. Normalitat: la distribució de les dades obtingudes segueix una distribució normal.

3. Homogeneitat de variàncies i covariàncies (esfericitat): les mostres obtingudes per a cada algorisme cal que provinguin de distribucions amb igual variança, i les covariàncies entre elles han de ser també iguals.

Si no es compleixen algunes d'aquestes condicions, caldrà estudiar quins efectes té sobre la validesa de l'anàlisi de variàncies. Si aquest efecte és prou important com per rebutjar la seva aplicació, s'haurà d'estudiar una alternativa que segurament passarà per utilitzar un test no-paramètric, malgrat implica reduir la informació que es tindrà en compte i, per tant, previsiblement tindran una menor potència (com ja s'ha comentat pel test de signes de Wilcoxon respecte el t-test, al capítol anterior).

La primera de les suposicions serà certa habitualment, en tant que es considera acomplerta si els problemes de prova no han estat triats per provocar un determinat resultat, sinó seguint els estàndards habituals: problemes del repositori UCI ([3]), problemes reals de domini mèdic, etc. En canvi, les altres dues suposicions caldrà que siguin comprobades sempre, molt en especial la tercerca, cosa que es fa en molt poques ocasions. No es coneix cap estudi més o menys exhaustiu, però un simple i breu anàlisi sobre els treballs presentats al *23rd International Conference on Machine Learning* (ICML2006), ens reafirma en aquesta consideració: de fet, en contades ocasions es va més enllà dels test per parelles.

La normalitat dels resultats obtinguts

La segona condició, sobre la normalitat de les dades, es presenta de la mateixa manera que amb el t-test, i per això no es repetiran els càlculs que són similars als ja fets a l'apartat 6.2.3. Només calen dos comentaris per remarcar algunes diferències. D'una banda, la condició s'ha de complir per totes les diferències entre els resultats obtinguts que es puguin comparar i, si ja era probable trobar violacions d'aquesta condició en el cas d'una comparació simple, ara encara ho serà molt més, per l'augment del nombre de comparacions per parelles que es poden realitzar.

D'altra banda, i en contraposició amb aquest darrer comentari, alguns autors com Hamilton ([116]) afirmen que és molt menys determinant per al bon funcionament de l'ANOVA que no pas ho era en el cas del t-test, especialment si només es viola normalitat en algunes de les diferències que es poden establir i, sobretot, si les distribucions obtingudes no mostren un fort comportament bi-modal.

En aquesta línia, alguns d'aquests autors continuen utilitzant l'anàlisi de variàncies malgrat no es compleixi la condició de normalitat (vigilant sempre la possible bi-modalitat de la distribució), emparant-se en dos factors: per una part, serà molt més determinant la tercera condició per al correcte funcionament d'aquest test; per altra, la inclusió dels test a posteriori, com es veurà immediatament a continuació, permet relaxar el compliment de les condicions per a l'ANOVA, especialment si es tracta d'un test a posteriori de tall conservador. Tal i com es veurà al final de l'apartat, la proposta que es fa en aquest treball va en aquesta línia, continuant l'estudi de les condicions excepte en el cas d'una clara violació de la normalitat, produïda per un comportament fortament bi-modal de la distribució dels resultats.

L'esfericitat i l'homogeneïtat de les variàncies

La tercera condició, anomenada d'esfericitat, pressuposa que les distribucions que donen lloc als resultats obtinguts per cada algorisme tenen els mateixos valors de les variàncies i, a més, que aquestes mesures obtingudes per cada algorisme estan incorrelades. És l'equivalent a l'homogeneïtat de variàncies que es demana per al t-test, tot i que pel fet que ara $M > 2$ apareix també l'efecte de la correlació entre els diferents algorismes. La dificultat de comprovar-la és bastant més gran que en aquell cas, fins el punt que l'opció proposada en aquest treball tendeix a desestimar-ne el càlcul exacte i s'opta per algunes aproximacions amb menor complicació matemàtica. Aquestes aproximacions, no obstant, aporten menys seguretat en les conclusions, com es veurà a continuació.

En primer lloc, i de manera breu, es mostren els principis que regeixen el càlcul de l'esfericitat, per un conjunt d'M algorismes aplicats sobre N problemes de prova. La majoria de les tècniques desenvolupades es basen en calcular el grau de semblança de la matriu de covariàncies calculada per als M algorismes respecte una matriu identitat o un múltiple d'aquesta. La matriu de covariàncies (MCV) ve definida per:

$$MCV_{ij} = Cov(X_i, X_j) \tag{7.10}$$

per als elements de fora de la diagonal ($i \neq j$), mentre que quan els índexs coincideixen es compleix:

$$MCV_i = Var(X_i) \tag{7.11}$$

Si la semblança és elevada es considera que es compleix la suposició d'esfericitat, perquè les variàncies seran similars entre elles i els resultats obtinguts pels algorismes estaran propers a la "no-correlació". Les dues tècniques habitualment utilitzades per al càlcul d'aquesta semblança són les aproximacions de Bartlett i de Mauchly ([7]), que comparen ambdues matrius (la de covariàncies i la identitat) a partir d'un càlcul de valors propis i determinants, i un test χ^2, amb un cert grau de significança α. Això implica que el propi test d'esfericitat té una certa probabilitat de cometre un error de "Tipus I", cosa fàcilment comprovable: l'aplicació d'aquests test sobre una simulació amb M mostres obtingudes d'una distribució normal mostra com, en un percentatge proper a α, els test preveuen que no es complexi la suposició d'esfericitat, és a dir, que les distribucions que han donat lloc a aquestes mostres són diferents.

El càlcul de l'esfericitat fet a partir d'aquests test té diversos inconvenients, que sovint no es tenen en compte: en la utilització de la metodologia exacta per a la comprovació de l'esfericitat, les condicions sota les quals es desenvolupa aquesta fa de nou difícil comprovar que s'està dins el domini d'aplicabilitat ([117]), amb la qual cosa s'arriba a un aparent bucle, doncs cal comprovar les condicions sota les quals es pot aplicar una metodologia, que serveix per comprovar que es compleixen les condicions per aplicar amb garanties l'anàlisi de variàncies.

A banda d'això, cal dir que els test d'esfericitats acostumen a ser molt restrictius: difícilment es poden trobar problemes reals en què es compleixin estrictament les condicions previstes per l'esfericitat, i malgrat tot l'anàlisi de variàncies continua essent una tècnica molt utilitzada. En el cas de l'exemple anterior, a partir de les dades publicades a [5], es troben els resultats expressats a la taula 7.3, on el valor p cal comparar-lo amb l'habitual $\alpha = 0.05$: si el resultat obtingut és menor, es pot rebutjar la hipòtesi de l'esfericitat. Com es pot veure, en cap dels tres casos estudiats es compleix. Aquest fet mostra com de restrictiu són aquests tests, i això ens porta a cercar alternatives que no descartin l'ús de l'anàlisi de variàncies si, malgrat no complir estrictament la condició d'esfericitat, es compleixen altres condicions.

De fet, anteriorment alguns autors han presentat diverses aproximacions a aquests càlculs. La més habitual és la proposada per Myers i Well ([8]), basant-se en el que anomenen la simetria composta, condició sobre el con-

	CBR	**SOMCBR-vot**	**SOMCBR-per**
χ^2	68.9	78.7	40.4
df	3	3	3
p	<0.001	<0.001	<0.001

Taula 7.3: Anàlisi de la suposició d'esfericitat per les dades obtingudes per cada una de les tres estratègies assajades a [5], els resultats de les quals es mostren a la taula 7.1. Com mostren els valors obtinguts per p, en els tres casos es poden rebutjar les hipòtesi nul·les sobre el compliment de l'esfericitat per un $\alpha = 0.05$. El valor χ^2 indica el retorn del test de comparació, i *df* els graus de llibertat del problema. El test que s'aplica per obtenir el resultat és el de Bartlett.

junt de dades i les seves variàncies i covariàncies que, tot i no ser condició necessària per a l'esfericitat, sí que n'és suficient.

En la comprovació d'aquestes condicions, es realitza un test d'homogeneïtat de variàncies com amb el t-test (sobre la de menor i major valor), però la comprovació sobre les covariàncies és tan complexa com la pròpia condició d'esfericitat, i sovint els càlculs es queden en una simple "inspecció visual": és el que en aquest treball es defineix com la comprovació de la *simetria composta dèbil* (scd). A la taula 7.4 es poden veure els resultats per als mateixos casos que abans, remarcant de nou el fet que es tracta d'una condició suficient, que no necessària: només en un dels tres casos es compleix la simetria composta dèbil i, per tant, es podria treballar amb l'anàlisi de variàncies, segons l'aproximació plantejada per Myers i Well, i la definició de scd que s'ha realitzat.

Protocol per a l'aplicació de l'ANOVA

Tot i això, en combinació amb les anteriors metodologies exposades, és possible elaborar un protocol per a la utilització de l'anàlisi de variàncies, que esquemàticament es mostra a la figura 7.1. La proposta comença amb el càlcul de l'estadístic F pel problema en qüestió i, d'acord amb els graus de llibertat associats, del valor crític F_α. Si el valor de l'estadístic és ordres de magnitud menor que el valor crític ($F << F_\alpha$), es proposa considerar-ho suficient com per utilitzar l'anàlisi de variàncies per discutir sobre la hipòtesi nul·la. Si no és el cas, s'inicia la comprovació de la normalitat de les dades i de l'esfericitat, en aquets darrer cas de dues maneres diferents i consecutives.

Si les condicions de normalitat es violen de manera "clara" (és a dir, la distribució és molt similar a una bi-modal), n'hi ha prou com per descartar

	CBR	**SOMCBR-vot**	**SOMCBR-per**
Var_{MAX}	80.31	67.5	66.02
Var_{min}	67.78	15.17	33.10
df	11	11	11
t	1.82	5.30	4.96
$t_{.05}$	2.20	2.20	2.20
Cov_{12}	1.09e4	9.54e3	7.63e3
Cov_{13}	1.05e4	-1.57e3	5.15e3
Cov_{23}	0.99e4	-1.60e3	5.62e3
scd	Si	No	No

Taula 7.4: Anàlisi del compliment (Si/No) de la suposició de simetria composta dèbil (scd) per les dades obtingudes per cada una de les tres estratègies assajades a [5]. Només en el primer cas es pot afirmar que es compleix.

l'anàlisi de variàncies i buscar una alternativa. En canvi, si no es compleix aquesta condició es pot continuar amb l'anàlisi de l'esfericitat, a través d'un dels test exposats anteriorment. Com aquests test són extremadament restrictius, es proposa donar una "segona oportunitat" a l'anàlisi de variàncies, i testejar el compliment de l'anomenada simetria composta dèbil (scd). Si es compleixen una de les dues condicions es proposa d'utilitzar l'anàlisi de variàncies. En cas contrari (tot i que la scd és suficient però no necessària per a l'esfericitat), s'optarà per una tècnica alternativa a l'anàlisi de variàncies.

Aquest protocol permet utilitzar l'anàlisi de variàncies en el cas de l'estratègia $SOMCBR-per$, doncs clarament $F >> F_\alpha$ (veure la taula 7.2). En els altres dos casos, en canvi, cal seguir l'"itinerari" marcat com es mostra gràficament a la figura 7.2: mentre en el cas del CBR és possible utilitzar l'anàlisi de variàncies (doncs es compleeix la simetria composta dèbil), en el cas del $SOMCBR-vot$ cal cercar una alternativa.

Cal esmentar en aquest punt que el que aquí es proposa és diferent del que en darrers treballs publicats es defensa (com a l'article de Demsar, [9], entre d'altres): en aquests casos s'afirma que la dificultat per a la comprovació del compliment de les condicions pel domini d'ús de l'ANOVA, i la pròpia dificultat de que aquestes es compleixin, justifica que directament s'opti per un cas no-paramètric. La proposta que es fa en aquest treball és, com a mínim, tenir en compte aquest protocol, i valorar en cada cas els punts a favor i en contra de la seva aplicació, doncs no convé oblidar que la utilització directe d'un test no-paramètric condicionaria el resultat, perquè s'estaria utilitzant menor informació que la que està disponible, i existeix major probabilitat de

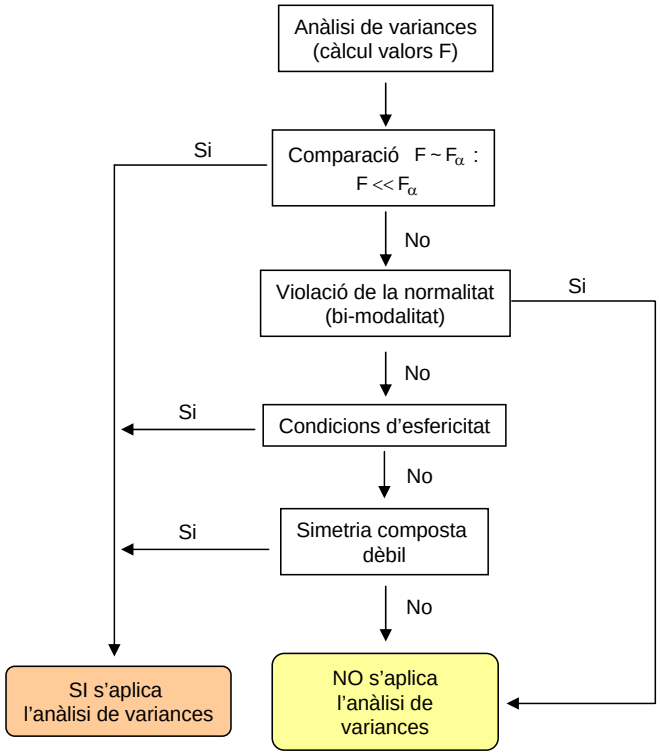

Figura 7.1: Protocol per a la utilització de l'anàlisi de variàncies, a partir de la comprovació de tot un conjunt de condicions exposades al text. En tot moment es considera que la mostra de N problemes de prova ha estat seleccionada a l'atzar d'entre la població de problemes de prova existents.

no detectar diferències significatives existents.

En cas que no sigui possible l'aplicació de l'anàlisi de variàncies, són possibles tot un conjunt d'alternatives, a triar en funció dels requeriments del problema:

- En primer lloc, el més habitual és fer el mateix que amb el t-test: buscar una alternativa no-paramètrica que, utilitzant menys informació dels resultats, permeti reduir les restriccions per a la seva aplicabilitat. Serà l'opció triada en els nostres problemes, i s'exposa àmpliament en l'apartat 7.4.

- Una altra opció és estalviar-se aquestes comprovacions és aplicar l'anàlisi de variàncies amb un valor d'α_{FW} menor que l'habitual. Té l'inconvenient que podria créixer la probabilitat de cometre un error "Tipus II",

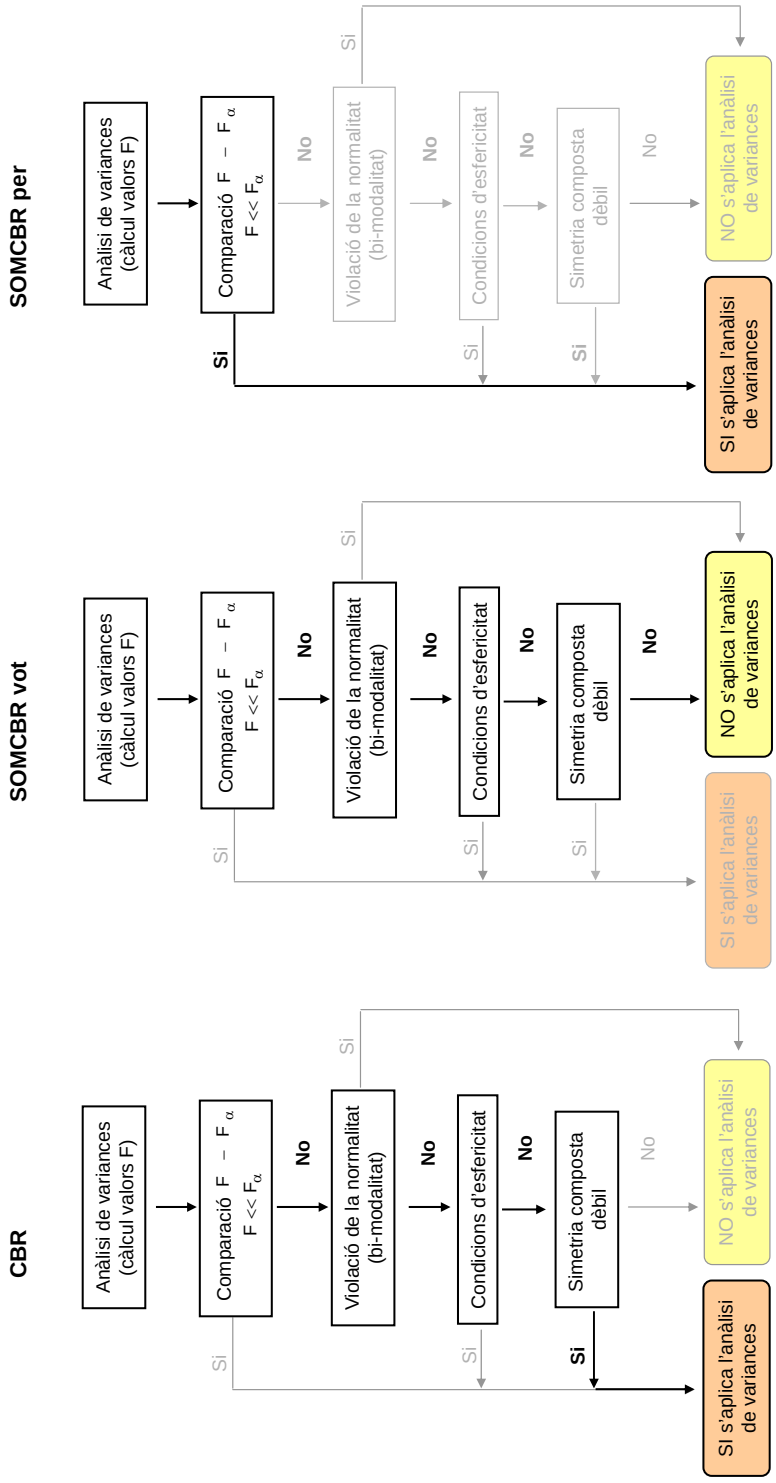

Figura 7.2: Aplicació del protocol descrit sobre les tres estratègies estudiades.

però pot ser una bona opció per assegurar-se el rebuig de H_0 i el passi als test a posteriori. És la preferida per alguns autors, tot continuant amb un test a posteriori de tall més conservador, en cas d'haver rebutjat H_0. Si el rebuig és un error fruit de l'incompliment de les condicions d'aplicabilitat de l'ANOVA, augmentar les restriccions del test a posteriori facilita posar-ho de relleu i descartar cap diferència significativa entre els comportament dels algorismes. No obstant això, i tenint en compte que la disminució d'α_{FW} provocarà un previsible augment de la probabilitat de cometre un error "Tipus II", la recomanació final que es farà preferirà la primera opció.

- Una altra possibilitat, que habitualment no es considera per la tipologia de problemes que es tracten en aquest treball, és refer el problema i optar per test multivariants, com MANOVA ([114]). No sembla necessari, vistes aquestes les opcions anteriors, i fins i tot no sempre és possible, doncs no sempre es tenen dos o més variables que indiquin algun aspecte de la bondat de l'algorisme (un exemple de cas multivariant es tractarà breument a l'apartat 9.3).

Finalment, cal dir que també es poden introduir algunes variacions en l'anàlisi de variàncies per tal de compensar les violacions de les condicions d'ús ([118], [119]), però una alternativa no-paramètrica combinada convenientment amb un test a posteriori prou conservador ja fa possible obtenir un resultat altament fiable i, a més, sense augmentar gaire la complexitat matemàtica dels càlculs. Vistes les dificultats de la comprovació i del propi compliment de les condicions d'aplicació de l'anàlisi de variàncies, el protocol global que es proposarà al final d'aquest capítol optarà clarament per opcions més aviat conservadores.

7.3.3 Test a posteriori

Com ja s'ha comentat, si l'anàlisi de variàncies permet rebutjar H_0, sabem que no és cert que les distribucions sota les dades obtingudes per cada algorisme tenen la mateixa mitjana. Per tant, que no és cert que tots els algorismes tinguin la mateixa bondat o, dit amb més precisió, que hi ha com a mínim una diferència significativa entre totes les possibles comparacions que es podrien realitzar (tant per parelles com complexes). No obstant això, el rebuig de la hipòtesi nul·la global no dóna més informació sobre quina o quines són aquestes comparacions que el provoquen.

Per respondre a aquesta qüestió, cal realitzar un test que aporti més

informació sobre el problema, especialment en aquells casos en què no és evident quina és la comparació que duu a una diferència significativa (no seria el cas de, per exemple, l'estratègia $SOMCBR-per$ segons els resultats de la taula 7.1). Si la comparació és planificada a priori, l'habitual test que segueix a un rebuig de la hipòtesi nul·la per l'ANOVA és una repetició d'aquest mateix test sobre la comparació simple planificada, tant en el cas que sigui per parelles com complexa. Si no fos el cas, l'estratègia és similar però associant-hi un α_{FW} corregit i, per tant, menor.

De fet, quan $M = 2$ obtenim $df_{BA} = 1$ i $df_{res} = (N - 1)$. Aquest darrer valor és el mateix que el *df* utilitzat pel t-test: com era d'esperar, doncs ambdós són test paramètrics basats en els mateixos arguments sobre la variabilitat observada en el dades, es pot demostrar que el resultat que s'obté per un ANOVA amb $M = 2$ és equivalent que el que s'obtindria amb un t-test sobre les mateixes dades.

Habitualment, el normal és fer aquestes comparacions entre aquell parell d'algorismes que a priori es volien comparar ($\overline{X}_1$ vs $\overline{X}_2$, per exemple) o bé realitzar el que també es coneix com un contrast: comparar el conjunt d'una sèrie d'algorismes amb un del sol, que és el que anomenem algorismes de contrast o de control ($\overline{X}_1$ vs $(\overline{X}_2+\overline{X}_3)/2$, per exemple). El darrer cas es pot utilitzar quan l'objectiu és demostrar que les modificacions introduïdes a A_2 i A_3 milloren el seu comportament respecte A_1, per posar un exemple d'un cas que succeirà habitualment.

En aquesta línia, i respecte les dades mostrades per les estratègies de la taula 7.1, es pot intentar demostrar que l'augment del valor de k aporta un millor comportament a l'algorisme classificador que amb $k = 1$. L'apartat anterior ha mostrat com l'anàlisi de variàncies es pot aplicar pel CBR i el $SOMCBR-per$. En el primer d'aquests dos casos no es detecta una diferència significativa, però sí en el segon, que és sobre el qual interessarà comprovar els efectes de k (tot i que la diferència entre els resultats obtinguts en mitjana ja indica clarament l'existència de diferències significatives). D'entrada, definim la hipòtesi nul·la a discutir:

$$H_0 : \overline{X}_1 = \frac{\overline{X}_2 + \overline{X}_3}{2} \tag{7.12}$$

i a partir d'ella introduim els coeficients c_i, definits com els factors que multiplicarien cada $\overline{X_i}$ en un plantejament de la hipòtesi nul·la amb igualtat a zero. Aquest plantejament seria

$$H_0 : \overline{X}_1 - \frac{\overline{X}_2 + \overline{X}_3}{2} = 0 \tag{7.13}$$

i d'aquí

$$c_1 = 1 \tag{7.14}$$
$$c_2 = c_3 = \frac{-1}{2}$$

Amb aquests valors, el càlcul de l'estadístic F sobre aquest contrast es pot deduir a partir del plantejament general mostrat a l'apartat 7.3.1 amb la introducció dels coeficients c_i:

$$F_{con} = \frac{MS_con}{MS_res}$$
$$MS_{con} = \frac{SS_{con}}{df_con} \tag{7.15}$$
$$SS_{con} = \frac{N\left(\sum_{i=1}^{M} c_i \overline{X_i}\right)^2}{\sum_{i=1}^{M} c_i^2}$$

on els graus de llibertat del contrast són sempre $df_{con} = 1$, com en una comparació simple per parelles.

A la taula 7.5 es pot veure el resultat per l'estratègia assajada ($SOMCBR-per$, l'única amb rebuig previ d'H_0). Es veu com la diferència és clarament significativa (doncs l'estadístic F calculat supera en molt als valors crítics), i ens permet afirmar que l'augment del valor de k va en favor de la millora de la bondat del classificador, confirmant així les conclusions obtingudes a partir de les dades de la taula 7.2.

	SOMCBR-per
SS_{con}	1409.50
df_{con}	1
MS_{con}	1409.50
MS_{res}	6.26
Estadístic F	225.21
$F_{crit}(\alpha = .05)$	4.26
$F_{crit}(\alpha = .01)$	7.82

Taula 7.5: Anàlisi dels contrastos respecte $\overline{X}_1$ per l'estratègia assajada a [5] en què l'anàlisi de variàncies és utilitzable i el valor de F permet rebutjar la hipòtesi nul·la global. Els resultats obtinguts ens permeten afirmar que l'augment del valor de k millora, de manera significativa i respecte l'error de classificació, el comportament de l'algorisme classificador.

7.3.4 La distància crítica

Un altre enfocament possible per als test a posteriori, que a més permet un major marge de maniobra respecte l'ajust de α_{FW}, és el càlcul de la distància crítica (CD), definida com la mínima distància requerida entre les dues magnituds comparades ($\overline{X}_1$ i $\overline{X}_2$, per exemple) per tal que es pugui considerar que existeix diferència significativa entre ambdues, a un determinat nivell de significança α. Per fer-ho, existeixen tot un conjunt de metodologies que porten a resultats diferents, doncs són també diferents les suposicions que fan.

S'exposaran, a continuació i de manera breu, les particularitats i forma de calcular de cadascuna de les principals metodologies per al càlcul de CD, i finalment es proposarà una manera d'actuar davant la tria dels diferents valors obtinguts per CD, amb un cas d'exemple basat en les dades publicades a [5].

Test LSD de Fisher

La primera de les metodologies presentades es basa en el càlcul de múltiples t-test després d'haver aplicat l'anàlisi de les variàncies. En tant que no ajusta el valor de α_{FW}, tal i com s'ha exposat a l'apartat 7.2, és el mètode que obté un menor valor de CD: amb més facilitat trobaria diferències significatives entre parelles, però també amb més facilitat cometria errors de "Tipus I" en les comparacions. L'ús és correcte per aquelles comparacions planificades, en el sentit definit abans, però si s'utilitza sobre un conjunt massiu de comparacions no planificades, el resultat de CD serà prou petit com per assegurar un elevat risc de cometre error "Tipus I".

El valor de CD ve expressat per:

$$CD_{LSD} = \sqrt{F_{1,res}} \sqrt{\frac{\sum_{i=1}^{M} c_i^2 MS_{res}}{N}} \tag{7.16}$$

on $F_{1,res}$ és el valor de l'estadístic F pel valor de α_{FW} considerat, i els c_i han estat definits a l'apartat anterior. El resultat que s'obté per CD_{LSD} és equivalent al que s'obtindria amb el t-test, doncs es pot demostrar l'equivalència entre els càlculs dels estadístics corresponents. Per exemple, s'observa aquest fet en les dades exposades per primer cop a la taula 7.1, per les quals es mostra el càlcul del CD_{LSD} en el contrast de $\overline{X}_1$ respecte $(\overline{X}_2 + \overline{X}_3)/2$, tal i com es veu a la taula 7.6. S'hi observa com el resultat sempre és coherent

amb l'obtingut per l'anàlisi de variàncies (taules 7.4 i 7.5): en tots els casos es rebutja la hipòtesi que enuncia igualtat de comportament entre algorismes.

	SOMCBR-per
F_{con}	225.21
$F_{crit}(\alpha = .05)$	4.26
Rebuig H_0 $(\alpha = .05))$	Si
$F_{crit}(\alpha = .01)$	7.82
Rebuig H_0 $(\alpha = .01))$	Si
CD_{con}	12.75
$CD_{LSD}(\alpha = .05)$	1.75
Rebuig H_0 $(\alpha = .05))$	Si
$CD_{LSD}(\alpha = .01)$	2.38
Rebuig H_0 $(\alpha = .01))$	Si

Taula 7.6: Càlcul de CD per la metodologia LSD, per l'estratègia assajada a [5] en què l'anàlisi de variàncies és utilitzable i el valor de F permet rebutjar la hipòtesi nul·la. Es realitza el contrast amb A_1, comparant per tant els valors de $\overline{X}_1$ amb $(\overline{X}_2 + \overline{X}_3)/2$. Els resultats obtinguts mostren l'equivalència d'aquest test amb l'obtingut per l'anàlisi de variàncies.

Test de Bonferroni-Dunn

Basat en la desigualtat de Bonferroni ([91]) i plantejat originalment per Dunn ([120]), és idèntic al test de Fisher descrit anteriorment, amb la diferència que ara es controla el valor de α_{FW} segons la relació ja mostrada anteriorment (equació 7.2), que es pot aproximar en general com

$$\alpha_{PC} \approx \frac{\alpha_{FW}}{c} \tag{7.17}$$

D'acord amb això, el càlcul de CD ve donat per l'expressió

$$CD_{B|D} = t_{B|D}\sqrt{\frac{\sum_{i=1}^{M} c_i^2 MS_{res}}{N}} \tag{7.18}$$

on $t_{B|D}$ és el valor crític de la distribució t al nivell de significança donat per α_{PC} i amb els graus de llibertat que determinen df_{res}. Degut al plantejament d'ajust de α_{FW}, el valor obtingut per CD és el major de tots, i per això diem que el test de Bonferroni-Dunn és el més conservador: la probabilitat

de cometre un error de "Tipus I" és mínima, però també és menor la capacitat de discriminar comportaments significativament diferents (i, per tant, major la possibilitat de cometre un error de "Tipus II").

Alguns autors ([114], [117]) han proposat modificacions si es considera que l'ajust de α_{FW} és massa sever. Per la nostra proposta de metodologia que es mostrarà a finals de l'apartat, ja va bé el resultat que s'obtindrà d'aquesta expressió.

Test HSD de Tukey

Aquesta versió del càlcul de CD ([121]) és la més recomanable quan el que es vol és fer totes les comparacions per parelles possibles (el que seria el cas extrem de les comparacions no planificades, [19]). Les inicials HSD provenen de *honestly significant difference*, i el valor de CD en aquesta ocasió ve donat per

$$CD_{HSD} = q_{M,df_{res}} \sqrt{\frac{MS_{res}}{N}} \tag{7.19}$$

on q segueix la distribució d'Student per rangs, tabulada en els textos d'estadística, en funció de la quantitat d'algorismes assajats (M) i els graus de llibertat df_{res}.

Si alguna de les suposicions de l'anàlisi de variàncies no es compleix, una versió modificada és el mètode de Tukey-Kramer ([106], [122]). D'altra banda, l'alternativa de Newman-Keuls ([123]) és també a vegades utilitzada, per bé que el valor obtingut és similar al donat pel test de Fisher ([113], [114]).

Test de Scheffé

Si el que es vol és mantenir fixat el valor de α_{FW}, independentment de quantes comparacions es facin a posteriori, aquesta és una molt bona opció, per bé que la pròpia condició converteix el test en força conservador, fins al punt que alguns autors arriben a proposar utlitzar un valor de $\alpha_{FW} = 0.10$. Sigui com sigui, en aquest cas el CD es calcula com

$$CD_S = \sqrt{(M-1)F_{df_{BA},df_{res}}} \sqrt{\frac{\sum_{i=1}^{M} c_i^2 MS_{res}}{N}} \tag{7.20}$$

on F és el valor tabulat per al α_{FW} escollit, i els valors corresponents de df_{BA} i df_{res}.

Degut al plantejament que es fa, el valor obtingut sovint és similar al $CD_{B|D}$. De fet, l'estudi de les característiques de cada metodologia ens permet construir un esquema de relació entre els valors obtinguts de CD del següent tipus:

$$CD_{LSD} \simeq CD_{N|K} < CD_{HSD} < CD_S \simeq CD_{B|D} \tag{7.21}$$

De fet, les diferències entre alguns d'aquests valors són molt petites i depenen fortament del número de comparacions a posteriori que es volen fer, així com de quines d'aquestes són per parelles o complexes.

7.3.5 Proposta per al correcte ús de CD

Com s'ha vist, els diferents test permeten obtenir diferents conclusions aplicats al mateix problema, en funció de les suposicions que es facin i de la voluntat d'assumir el risc de cometre un error de "Tipus I", doncs els valors que s'obtenen de CD són diferents dependent d'aquestes suposicions. Davant d'això, a continuació es fa una proposta d'ús del valor de CD, a partir de les seves diferències i d'elements que es mostren a [19] i a [117].

A banda d'aquesta diferència en els valors obtinguts per cada metodologia, cal dir també que alguns dels test a posteriori presentats depenen en part del compliment de les condicions per a l'aplicació de l'ANOVA, en especial la condició d'esfericitat. La discussió sobre les modificacions a introduir per salvar la condició d'esfericitat ([106], [113], [114], [122]) depassa la voluntat d'aquest treball, però s'han de tenir presents a l'hora d'establir un mètode prudent de tractar aquests valors de CD. La proposta és ben simple en quant a l'execució: en primer lloc, i un cop rebutjada la hipòtesi H_0 global per l'anàlisi de variàncies, es calculen els valors CD_{LSD} i $CD_{B|D}$ per les comparacions que es vulguin dur a terme (que, segons l'expressat, seran els valors mínim i màxim que es poden obtenir, respectivament). Per cada comparació k es planteja la hipòtesi nul·la $H_{0,k}$ i s'obté la diferència $\overline{\Delta X}_k$, amb cada una de les quals es fa el següent plantejament:

1. Si $\overline{\Delta X}_k < CD_{LSD}$, no hi ha diferència significativa per la comparació k, i per tant no es pot rebutjar $H_{0,k}$.

2. Si $\overline{\Delta X}_k > CD_{B|D}$, sí que hi ha diferència significativa per la comparació k, i per tant es pot rebutjar $H_{0,k}$.

3. Si $CD_{LSD} < \overline{\Delta X}_k < CD_{B|D}$, no es pot treure una conclusió clara.

En el darrer cas, i donades les dificultats per comprovar el compliment de les condicions d'esfericitat i per l'ajust d'α_{FW}, la proposta és optar directament per un test a posteriori no-paramètric, que tindrà menys capacitat de detectar diferències significatives, però que aportarà major fiabilitat al resultat un cop s'ha rebutjat H_0.

Una altra opció per intentar resoldre la darrera possibilitat és recalcular els valors CD, amb una anàlisi de variàncies tenint present només els algorismes de la comparació k. Diferents autors ([19], [117]) demostren com d'aquesta manera el valor de MS_{res} varia, i l'interval obtingut entre CD_{LSD} i $CD_{B|D}$ és més fiable, reduint el valor de la diferència $CD_{LSD} - CD_{B|D}$. No obstant això, mai s'elimina la possibilitat de tenir un cas que compleixi la condició establerta a l'opció 3.

A la taula 7.7 es mostren els resultats sobre el contrast de $\overline{X}_1$ respecte $(\overline{X}_2+\overline{X}_3)/2$, i es veu com per l'estratègia $SOMCBR-per$ el valor de CD_{con} és major que el de ΔX, per la qual cosa s'està a l'opció 2, i es pot rebutjar l'opció nul·la del contrast: com ja s'havia vist, clarament l'augment de k millora el comportament de l'algorisme classificador.

	SOMCBR-per
ΔX	12.75
CD_{LSD}	1.75
$CD_{B\|D}$	2.34
Opció	$\overline{\Delta X} > CD_{B\|D}$
Rebuig H_0 ($\alpha = 0.05$))	Si

Taula 7.7: Aplicació de la proposta per al correct ús de CD, en el cas de l'estratègia assajada a [5] en què l'anàlisi de variàncies és utilitzable i el valor de F permet rebutjar la hipòtesi nul·la. En el contrast amb A_1, comparant per tant els valors de $\overline{X}_1$ amb $(\overline{X}_2 + \overline{X}_3)/2$, els resultats obtinguts asseguren el rebuig de la hipòtesi nul·la i, per tant, asseguren que l'augment de k millora el comportament de l'algorisme classificador.

7.4 Alternatives no paramètriques

A l'apartat anterior, s'ha discutit àmpliament la possibilitat que no es compleixin les condicions que permeten l'aplicació de l'anàlisi de variàncies per un conjunt d'M algorismes testejats sobre N problemes de prova. Una de les alternatives possibles si no es compleixen aquestes condicions (veure el protocol representat a l'esquema de la figura 7.1), com ja passava en el cas de les

comparacions simples (apartat 6.3), és optar per un test no-paramètric, sobre el qual es realitzen moltes menys suposicions. Un test d'aquest tipus utilitza menys informació sobre el problema i, per tant, tindrà menys capacitat de discriminar una diferència de comportament entre algorismes.

7.4.1 Test de Friedman

L'alternativa no paramètrica més habitual és l'anomenat test de Friedman ([6], [124], [125]). Aquest test és una extensió del test de signes binomial utilitzat per comparacions simples, millorant clarament les prestacions respecte la corresponent matriu de guanys discutida al final de l'apartat 6.3.3.

Les suposicions per aplicar el test de Friedman són menors que les necessàries per fer el mateix amb l'anàlisi de variàncies: tan sols cal que la mostra de N problemes de prova hagi estat seleccionada a l'atzar d'entre la població de problemes de prova existents, i que de les mesures disponibles es pugui treure informació ordinal, essent originalment una variable contínua ([109]). És a dir, que permeti ordenar-les de major a menor per cada problema de prova. Excepte condicions excepcionals, ambdues es compliran sempre per als problemes que es tracten en aquest treball.

De fet, aquestes condicions ja indiquen amb quines dades treballa aquest test, i quina n'és la principal conseqüència: dels valors obtingut de l'aplicació dels M algorismes sobre els N problemes de prova es passa a una magnitud que determina l'ordre, de major a menor, dels resultats obtinguts per a cada un dels N problemes de prova, de manera similar a com es feia a l'apartat 5.2. És evident, per tant, que es sacrifica informació respecte l'anàlisi de variàncies, que treballa directament amb el resultat obtingut, i per tant la capacitat de determinar diferències significatives serà menor. Però el fet que les condicions per a la seva aplicabilitat siguin menors i, sobretot, molt més fàcil de complir-se, augmenta la fiabilitat del resultat i permet l'aplicació del test sense els càlculs necessaris per confirmar el domini d'ús de l'anàlisi de variàncies.

El plantejament, doncs, és bastant similar al fet per al test de Wilcoxon respecte el t-test. A més, el mateix Friedman va mostrar al seu moment com no hi ha una gran diferència entre els resultats obtinguts en la majoria dels casos: a la taula 7.8 es transcriuen els resultats que va obtenir a [6], després d'aplicar experimentalment ambdues tècniques d'anàlisi a 56 problemes independents. Malgrat teòricament hauria d'existir una important diferència (si les condicions d'esfericitat es compleixen, l'ANOVA és més capaç de determinar diferències significatives i, en canvi, quan no és el cas no es pot

aplicar perquè pot dur a conclusions errònies), en la citada taula es veu com en pràcticament tots els càlculs els resultats són coherents per un i altre mètode. Només en 6 dels casos (destacats en negreta) un dels mètodes troba significant una diferència i l'altre no ho fa, mentre que en la resta el comportament és similar: quan un troba diferència significativa per $\alpha = 0.01$, l'altre test ho troba per $\alpha = 0.05$, com a mínim.

		ANOVA		
		$p < 0.01$	$0.01 < p < 0.05$	$p > 0.05$
Friedman	$p < 0.01$	16	1	0
	$0.01 < p < 0.05$	4	1	**4**
	$p > 0.05$	0	**2**	28

Taula 7.8: Resultats originals publicats a [6], on es mostren els resultats de l'aplicació de l'anàlisi de variàncies i del test de Friedman sobre 56 problemes independents. Els valors indiquen en quantes ocasions els corresponents anàlisis retornen un valor $p < 0.01$, $0.01 < p < 0.05$ o $p > 0.05$.

El càlcul dels estadístics de Friefman

La metodologia per a l'aplicació del test de Friedman és com segueix: en primer lloc, i per cada problema de prova, s'ordenen els diferents algorismes segons el resultat de precisió obtingut, de major (1) a menor (M), establint el que s'anomenen els rangs corresponents a cada algorisme i per al problema de prova j, R_{ij}. En cas que amb els algorismes i_1 i i_2 s'obtingui el mateix resultat sobre el problema de prova j ($X_{i_1 j} = X_{i_2 j}$), i per tant es dubti entre atorgar-los una valor R o $R + 1$, s'atorga a ambdós el valor mig d'aquests, $R + 1/2$. Si hi ha més de dos algorismes amb igual resultat sobre un mateix problema, s'opera extrapolant aquesta operativa. És l'únic cas pel qual es permet treballar amb rangs no naturals.

A partir d'aquí, s'obtenen els valors mitjans dels rangs per a cada algorisme, $\overline{R}_i$, i a partir d'un test χ^2 es comparen aquests resultats amb el valor esperat en cas que el comportament de tots els algorismes sigui equivalent: si realment els M algorismes tinguessin un comportament equivalent, el rang del resultat sobre cada problema de prova seria una variable aleatòria amb distribució uniforme i valors possibles entre 1 i M. Per tant, el valor esperat de $\overline{R}_i$ seria $(M + 1)/2$ per a cada algorisme.

El test de Friedman, per tant, avalua la probabilitat que les desviacions en els rangs mitjos calculats respecte el previst si tots els algorismes tinguessin

un comportament equivalent siguin fruit de l'atzar, o bé del diferent comportament d'aquests algorismes. Si considerem, com habitualment, M algorismes sobre N problemes de prova, l'estadístic de Friedman és igual a:

$$\chi_F^2 = \frac{12N}{M(M+1)} \left(\sum_{j=1}^{M} R_j^2 - \frac{M(M+1)^2}{4} \right) \tag{7.22}$$

Aquest estadístic es compara amb la distribució χ^2 per $(M-1)$ graus de llibertat, tabulada a qualsevol text estadístic de referència. Força autors ([108], [109]) recomanen ajustar aquest valor si, a l'hora de determinar els rangs, apareixen resultats no naturals. Així, si considerem t_j com el nombre de resultats en què hi ha empat entre dos o més algorismes per al problema de prova j, es pot calcular un factor corrector C que és igual a

$$C = 1 - \frac{\sum_{j=1}^{N}(t_j^3 - t_j)}{N(M^3 - M)} \tag{7.23}$$

i que s'aplica dividint l'estadístic χ_F^2 per obtenir l'estadístic corregit $\chi_{F,cor}^2$:

$$\chi_{F,cor}^2 = \frac{\chi_F^2}{C} \tag{7.24}$$

Posteriorment al treball de Friedman, Iman ([126]) desenvolupà un nou estadístic, amb major capacitat d'explicitar diferències significatives, que es calcula a partir de l'estadístic de Friedman:

$$F_I = \frac{(N-1)\chi_F^2}{N(M-1) - \chi_F^2} \tag{7.25}$$

i es compara amb la distribució F amb $(M-1)$ i $(M-1)(N-1))$ graus de llibertat. Els valors de les distribucions χ^2 i F que s'utilitzen per realitzar aquests test són només aproximacions dels valors crítics reals dels estadístics, vàlids per valors grans de N i M (habitualment, per $M > 5$ i $N > 10$). En cas contrari, cal calcular els valors crítics de cada estadístic per mètode exactes ([127]).

Aquestes dues estratègies són les més habituals per als test múltiples no-paramètrics, donada la facilitat d'aplicació i els bons resultats que habitualment s'obtenen. Altres tècniques es poden consultar a [110], [128] o [129], però no aporten millores determinants als problemes que es tracten en aquest treball.

Dos exemples

Un primer exemple d'aplicació del test de Friedman es pot veure a la taula 7.9, on es mostren els valors dels rangs per a cadascun dels problemes de prova testejats a l'article de Bacardit i Garrell ([4]). D'entrada, es pot veure l'efecte de treballar amb informació ordinal en el fet que les diferències observades pels valors mitjos de la mesura de bondat no tenen perquè mantenir-se en els valors de $\overline{R}_i$. Per exemple, el valor de la precisió sembla indicar que l'algorisme C4.5 té un millor comportament respecte l'algorisme IB1 (78.77% respecte 78.65%), però en canvi el rang de IB1 és bastant més elevat que el de C4.5 (4.37 respecte 5.47). Aquest exemple mostra com la utilització dels rangs converteix el test no-paramètric en molt més robust davant la presència d'*outliers*.

A la taula 7.10 es veuen els resultats obtinguts d'aplicar les fórmules definides anteriorment sobre els resultats mostrats a la taula 7.9. Aquests resultats, a banda d'il·lustrar pròpiament l'aplicació del test de Friedman, mostren les limitacions dels test destinats a la comparació simple un cop extrapolats a la comparació múltiple: els valors obtinguts dels estadístics mostren com no hi ha cap diferència significativa entre els vuit algorismes testejats sobre quinze problemes de prova. En canvi, al final de l'apartat 6.3.3, semblava que la matriu de guanys construida a partir del test binomial ens permetia afirmar que hi havia diferències significatives entre alguns d'aquests algorismes.

La conclusió és clara: l'anàlisi per parelles ($M = 2$) dels resultats obtinguts en un problema de comparació múltiple ($M > 2$) pot portar a conclusions errònies. En canvi, un test preparat per comparacions múltiples detecta l'existència de diferències significatives de manera molt més fiable, ja sigui a partir d'aproximacions paramètriques (l'anàlisi de variàncies, si es pot utilitzar) o no-paramètriques (com aquest cas, amb el test de Friedman). Si aquest darrer hagués permès el rebuig de la hipòtesi nul·la caldria analitzar aleshores quina o quines comparacions simples ho provocarien. A l'apartat **??**, s'estudiarà a fons aquest mateix problema.

Un altre exemple d'interès és el resultat que s'obté de l'aplicació del test de Friedman sobre la segona de les estratègies assajades a [5], $SOMCBR-vot$, amb els resultats que es mostren a la taula 7.1. Recordem que anteriorment s'havia mostrat com no es compleixen les condicions suficients com per aplicar l'anàlisi de variàncies i, per tant, l'alternativa seria l'assaig d'un test no-paramètric. Si s'aplica el test de Friedman sobre el citat problema, s'obtenen els resultats que es mostren a la taula 7.11: el test no-paramètric no permet

Dataset	ADI		ADI1		ADI2		ADI3		ADI4		ADI5		C4.5		IB1	
	%	R	%	R	%	R	%	R	%	R	%	R	%	R	%	R
bpa	63.7	4.5	63.7	4.5	63.9	3	62.6	8	63.3	6	63.2	7	68.4	1	64.5	2
bps	80.6	4.5	80.7	2.5	80.7	2.5	79.6	8	80.6	4.5	80.0	7	80.1	6	83.2	1
bre	95.6	7	95.8	4.5	96.0	1.5	95.7	6	95.8	4.5	95.9	3	95.4	8	96.0	1.5
gls	66.4	6	66.5	4.5	67.9	1.5	67.9	1.5	67.8	3	66.5	4.5	65.8	8	66.3	7
h-s	80.4	4	80.2	5	80.7	1	79.8	6	80.5	3	80.6	2	76.3	7	74.1	8
ion	91.6	5	90.9	6	92.2	2	91.7	4	92.7	1	92.0	3	89.8	7	86.9	8
lrn	68.1	5	68.0	6	68.6	3.5	67.8	7	68.9	2	69.0	1	68.6	3.5	61.4	8
mmg	65.0	6	66.1	4	66.0	5	67.8	1.5	67.0	3	67.8	1.5	64.8	7	63.5	8
pim	74.4	4.5	75.1	2	74.7	3	74.3	6	75.3	1	74.4	4.5	73.1	7	70.3	8
son	74.6	2	73.2	5	73.1	6	72.3	7	74.3	3	73.5	4	71.5	8	87.3	1
thy	91.9	5	92.0	3.5	91.4	8	91.6	6	92.0	3.5	91.5	7	92.6	2	96.8	1
veh	66.0	7	66.4	5	66.1	6	65.6	8	66.7	3	66.5	4	73.6	1	69.4	2
wdbc	93.8	4.5	93.7	6	93.8	4.5	93.9	3	94.0	2	93.7	7.5	93.7	7.5	95.6	1
wine	92.7	4	92.5	6	92.2	7.5	92.6	5	93.0	3	92.2	7.5	94.1	2	95.6	1
wpbc	75.7	5	75.5	6	76.1	3	75.9	4	77.1	1	76.8	2	73.7	7	68.8	8
Mitjana	78.7	4.9	78.7	4.7	78.9	3.9	78.6	5.4	79.3	2.9	78.9	4.4	78.8	5.5	78.6	4.4

Taula 7.9: Resultats de [4], amb l'aplicació de les cinc variants de l'algorisme ADI assajades, el propi algorisme ADI i els algorismes $C4.5$ i $IB1$. S'hi han afegit els rangs, tal i com s'han definit en el text.

	$M = 8,\ N = 15$
χ^2_F	12.42
C	0.96
$\chi^2_{F,cor}$	12.89
df	7
$\chi^2_{F,.05}$	14.07
Rebuig H_0	No
F_I	1.88
df	7 x 98
$F_{I,.05}$	2.10
Rebuig H_0	No

Taula 7.10: Resultats de l'aplicació del test de Friedman sobre els resultats de la taula 7.9. Els valors de l'estadístic obtinguts no permeten rebutjar l'opció nul·la.

rebutjar la hipòtesi nul·la, malgrat estar molt a prop del valor crític. Aquest resultat és una prova més de la poca capacitat que el test de Friedman té per posar de manifest diferències significatives, més a prop del que seria el test binomial que no pas el de Wilcoxon, si es fa una comparativa amb les comparacions simples ([19]).

	$M = 3,\ N = 13$
χ^2_F	5.69
C	1
$\chi^2_{F,cor}$	5.69
df	2
$\chi^2_{F,.05}$	5.99
Rebuig H_0	No
F_I	3.36
df	3 x 13
$F_{I,.05}$	3.40
Rebuig H_0	No

Taula 7.11: Resultats de l'aplicació del test de Friedman sobre els resultats de la segona estratègia introduida a [5], els resultats de la qual es mostren a la taula 7.1. Els valors de l'estadístic obtinguts no permeten rebutjar l'opció nul·la.

Aquest cas mostra l'extrem a què ens duu l'operativa proposada a l'apartat 7.3.2: les condicions no són suficients per aplicar l'anàlisi de variàncies, que donada la major capacitat discriminant permetria rebutjar H_0, i el test no-paramètric no té prou capacitat per fer-ho, malgrat situar-se en valors de F_I propers al crític $F_{I,.05}$.

Alguns autors, en aquest cas, proposarien el rebuig d'H_0 i l'aplicació del test a posteriori (segurament en relació amb la cita que encapçala aquest capítol), però la tesi d'aquest treball és més conservadora, i mantindria la reticència al rebuig d'H_0. Si fos el cas que *realment* existís una diferència significativa com per rebutjar H_0, de ben segur apareixeria ràpidament en un test no-paramètric tot ampliant una mica la col·lecció de problemes de prova utilitzada.

7.4.2 Test a posteriori

De manera similar a l'exposat en l'apartat de l'anàlisi paramètric, en cas que el test múltiple ens permeti rebutjar la hipòtesi nul·la (i només en aquest cas), cal aplicar un test a posteriori per estudiar quin o quins dels algorismes tenen comportaments diferents, a partir de les dades obtingudes sobre els problemes de prova de l'assaig. El rebuig d'H_0 afirma que existeix, com a mínim, una diferència de comportament significativa al nivell α utilitzat (per parelles o complexa), i els test a posteriori ens han de permetre determinar

quina o quines són.

Els mètodes que aquí es presentaran es basen en el càlcul de la distància crítica, CD, definida ja anteriorment com mínima distància requerida entre les dues magnituds comparades per tal que es pugui considerar que existeix diferència significativa entre ambdues, a un determinat nivell de significança α.

Depenent del control que s'efectuï sobre α_{FW} apareixeran diversos resultats, que s'analitzaran en funció del tipus de pregunta que es vulgui respondre: uns plantejaments estan pensats per quan es volen comparar tot un conjunt d'algorismes respecte a una altre que anomenarem control (*planned comparisons*), mentre que uns altres no predeterminen un algorisme respecte el qual comparar-se (*unplanned comparisons*), i per tant cerquen diferències significatives en totes les comparacions possibles. Si és el primer cas, s'exposaran altres metodologies no basades directament en el càlcul de CD, donada la capacitat de discriminar diferències significatives sense perdre precisió en el resultat.

Cal dir també que no es pot assegurar que aquests test tinguin prou capacitat com per determinar quina o quines comparacions duen al rebuig de la hipòtesi nul·la: podria passar que, malgrat Friedman permetés rebutjar l'H_0 global, cap d'aquests test detectés cap diferència significativa. Això voldria dir que, o bé aquest test no s'ha assajat sobre la diferència (previsiblement complexa) que permet el rebuig d'H_0, o bé simplement que aquests test no tenen prou capacitat com per detectar-les.

Test de Nemenyi

Si l'anàlisi de Friedman permet rebutjar la hipòtesi nul·la, i no es disposa de cap algorisme de control, el test de Nemenyi ([130]) aporta una mesura de tall conservador per a la distància crítica, tenint en compte que es calcula sobre totes les possibles $M(M-1)/2$ comparacions per parelles (amb el corresponent efecte sobre l'adjust de α_{FW} i α_{PC}).

En aquest cas, aquesta distància crítica es calcula a partir de la següent expressió:

$$CD = q_\alpha \sqrt{\frac{M(M+1)}{6N}} \tag{7.26}$$

on M és el número d'algorismes assajats sobre N problemes de prova, i q és l'estadístic de la ditribució d'Student per rangs convenientment normalitzat ([9], [19]). El valor trobat és la diferència mínima entre els rangs $\overline{R_1}$ i $\overline{R_2}$ que

ens permet assegurar que dos algorismes A_1 i A_2 mostren comportaments diferents.

Test de Bonferroni-Dunn

També en el cas que l'anàlisi de Friedman permeti rebutjar la hipòtesi nul·la, sovint és d'interès realitzar la comparació respecte un algorisme de control, habitualment aquell respecte el qual es compara la millora introduïda en el treball que es presenta. En aquest cas, el test de Bonferroni-Dunn ([120]) controla el valor d'α_{FW} dividint-lo per les $(M-1)$ comparacions que es realitzaran, i es calcula amb la mateixa expressió que en el cas del test de Nemenyi, però ajustant q_α a aquest nou número de comparacions.

Degut al diferent ajust d'α_{FW}, el test de Bonferroni-Dunn tindrà sempre més capacitat de determinar diferències significatives: això vol dir que no té sentit fer totes les comparacions per parelles possibles, si el que interessa és la comparació amb un algorisme de control. En cas contrari, es detectaria un número menor de diferències significatives al que realment existeix.

Test de Holm

Les dues metodologies exposades fins ara per a realitzar test a posteriori es basen en el càlcul d'un valor de la distància crítica (CD), i la discussió posterior dels resultats obtinguts pels rangs a partir d'aquest valor. El test de Holm ([131]), en canvi, utilitza el que es coneix com les metodologies de múltiples passos (*step-up* o *step-down*, segons el cas), que permet un anàlisi respecte l'algorisme de control de manera més precisa que el test de Bonferroni-Dunn, i sense augmentar les restriccions sobre el problema.

En primer lloc, cal calcular l'estadístic per a cada algorisme i, ordenat de pitjor a millor comportament respecte la seva comparació amb l'algorisme de control, que suposarem té un rang mig R_0. Així, al pitjor algorisme li correspondrà l'estadístic z_1, al segon pitjor z_2, etc. Els valors es calculen a partir de

$$z_i = \frac{(R_i - R_0)}{\sqrt{\frac{M(M+1)}{6N}}} \tag{7.27}$$

i d'aquí s'estableix el valor p_i aproximant a partir d'una distribució normal. Degut a l'ordenació que es planteja, sempre es complirà que $p_1 \leq p_2 \leq ... \leq$

p_{M-1}. A continuació, es comparen els valors de p_i amb el valor

$$\alpha_i = \frac{\alpha_{FW}}{M - i} \tag{7.28}$$

i es considera que es podran rebutjar les hipòtesis nul·les per les comparacions de l'algorisme de control amb aquells algorismes en que el valor obtingut de p_i sigui menor que α_i. El mètode de Holm es considera un mètode de pas descendent, doncs l'anàlisi es pot començar sempre per l'algorisme amb una diferència de rang més elevada respecte l'algorisme de control (estadístic z_1), i per tant el valor de p més significatiu (p_1), i anar descendint en la significança del resultat fins arribar al primer cas en què $p_i > \alpha_i$: arribat a aquest punt ja no cal continuar, doncs no es podran rebutjar cap de les restants comparacions.

Com a complement del test de Holm, Hommel ([132]) i Hochberg ([133]) han desenvolupat també estratègies similars, però que no aporten grans diferències en el resultat ([134]). Per aquest motiu, el test de Holm és el més aconsellable d'aplicar en un problema en què l'anàlisi de Friedman hagi permès rebutjar H_0, i en què interessi realitzar la comparació respecte un algorisme de control.

7.4.3 Aplicació i capacitat de rebuig en els test a posteriori

Un bon exemple per a l'estudi de les potencialitats de cada una d'aquestes tècniques són els resultats publicats a [2], en què es poden estudiar els resultats obtinguts en funció de la complexitat dels diferents problemes de prova (un total de 56), sobre els quals s'assagen fins a 13 diferents estratègies de clusterització de la memòria de casos, incloent-hi la metodologia CBR "clàssica" (és a dir, considerant tots els elements que formen aquesta memòria). Posteriorment, a l'apartat **??**, s'estudiarà a fons aquest problema.

En aquest cas presentarem l'estudi fet sobre els problemes de prova "tipus B" (veure l'apartat 4.7), els resultats dels quals es mostren a la taula 7.12, pel que fa als rangs, i a la taula 7.13 pel que fa als resultats en percentatge d'error.

A la taula 7.14 es mostren a quina estratègia correspon cada un dels algorismes analitzats. Basant-se en la nomenclatura explicitada en el seu moment, S indica que la memòria de casos s'ha clusteritzat a partir d'un procediment SOM, mentre que OBM, EBM, $PEBM$ i OAN indiquen diferents estratègies en la recuperació (*Only the Best Model*, *Elements from*

Dataset	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}
biopsia	1	7	4	10	11	5	2	6	3	8	12	9	13
ddsm2c1	7.5	2	7.5	13	12	9	1	5.5	4	5.5	3	11	10
ddsm2c4	8	7	5	10	13	6	2	3	4	9	12	1	11
glass2c3	13	12	10	7.5	9	1.5	3.5	1.5	3.5	6	7.5	5	11
glass2c5	1	3	2	8	11	6	4	7	5	10	13	9	12
glass2c6	1.5	3	1.5	9	11	4.5	4.5	7	6	10	12	8	13
hepatitis	13	2	8	6	7	10	12	9	11	1	4	3	5
ionosphere	3	1	4	8	13	6	2	7	5	10	11	9	12
mias3c2c1	1	7	2.5	11.5	8	6	2.5	5	4	10	13	9	11.5
mias3c2c2	7	3	1	9	11	2	5	6	4	8	12	13	10
mias3c2c3	1	5	3	11	7	9.5	4	6	2	13	8	12	9.5
miasbi2c1	1	13	4	11.5	9	5	2	6	3	11.5	7	10	8
miasbi2c2	1	4	5	10	8.5	8.5	2	12	3	13	6	11	7
miasbi2c3	3	11	4	12	7	6	1	10	2	9	5	13	8
segment2c3	1	7	4	8	13	6	2	5	3	9	10	11.5	11.5
segment2c4	2	11	5	10	12	7	3	8	4	1	9	13	6
segment2c5	1	7	4	9	12	6	2	5	3	10	8	11	13
sonar	1	7	2	10	13	5	4	6	3	8	11	9	12
tao	3	1	5	9	12	6	2	8	4	10	7	13	11
thy2c3	1	3	2	7	13	5	4	8	6	9	10	11	12
vehicle2c1	1	10	4	7	13	5	3	6	2	9	11	8	12
vehicle2c2	3	8	2	7	12	5	4	6	1	10	11	9	13
vehicle2c3	2	7.5	4	11	10	7.5	3	5	1	6	13	12	9
vehicle2c4	1	9	4	7.5	13	5	2	6	3	10	7.5	11	12
wav2c1	4	11.5	7	11.5	13	9	5	8	6	1	10	2	3
wav2c2	2	6	4	9.5	12	5	1	7	3	8	9.5	11	13
wav2c3	2	8	1	9	11	5	4	6	3	7	12	10	13
wbcd	7	12	4	10	2	5	3	8	1	9	6	13	11
wisconsin	2	1	5	12	11	6	3	8	4	7	10	9	13
wpbc	1	13	9.5	4	5.5	9.5	3	11	7	12	2	5.5	8
Mitjana	3.2	6.7	4.3	9.3	10.5	6.1	3.2	6.7	3.8	8.3	9.1	9.4	10.5

Taula 7.12: Resultats per a l'aplicació de les estratègies desenvolupades a [2] sobre el conjunt de problemes de prova amb complexitat mitjana, segons la definició donada a 4.7. El valor que es mostra és el rang R_{ij}, tal i com ha estat definit.

Dataset	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}
biopsia	16.9	22.6	20.6	23.2	23.2	21.2	18.6	21.8	18.6	22.8	23.5	23.1	23.6
ddsm2c1	17.6	16.6	17.6	18.4	17.8	17.6	16.4	17.4	17.3	17.4	17.2	17.7	17.6
ddsm2c4	23.6	23.3	23.0	23.7	25.6	23.2	22.6	22.9	22.9	23.7	25.2	21.8	24.0
glass2c3	12.2	10.9	10.3	9.7	9.8	8.7	8.8	8.7	8.8	9.4	9.7	9.2	10.4
glass2c5	20.1	20.4	20.3	23.3	25.2	21.8	20.6	22.6	21.7	24.1	26.6	23.7	25.9
glass2c6	18.7	19.2	18.7	25.5	26.8	20.7	20.7	20.9	20.8	26.5	28.7	24.7	29.2
hepatitis	33.6	23.2	25.3	24.5	24.7	27.3	30.3	26.9	29.7	23.1	23.9	23.4	24.2
ionosphere	13.1	12.4	13.2	15.1	19.6	13.5	12.6	14.2	13.3	16.6	18.2	15.6	19.5
mias3c2c1	19.9	23.3	21.4	26.2	25.0	23.0	21.4	22.6	22.2	26.0	27.9	25.4	26.2
mias3c2c2	13.4	12.4	11.5	14.4	15.1	11.9	12.9	13.2	12.6	13.9	15.5	15.8	15.1
mias3c2c3	27.0	30.4	29.4	33.1	32.4	33.0	29.7	31.8	29.1	34.8	32.9	34.6	33.0
miasbi2c1	12.8	19.4	14.5	18.9	18.1	16.4	13.0	17.1	13.5	18.9	17.2	18.8	17.7
miasbi2c2	23.4	26.3	26.9	28.7	28.4	28.4	24.3	28.9	24.6	29.1	27.6	28.8	28.3
miasbi2c3	20.6	24.1	21.5	24.4	22.8	22.2	19.2	23.8	19.5	23.4	22.1	24.8	22.9
segment2c3	1.4	3.5	2.3	4.2	4.9	3.2	2.0	3.1	2.1	4.3	4.4	4.6	4.6
segment2c4	1.2	4.4	2.2	4.1	4.5	3.1	1.7	3.2	1.8	1.0	3.7	4.7	3.0
segment2c5	2.7	5.8	4.0	6.3	7.0	4.9	3.4	4.8	3.7	6.4	6.2	6.7	7.1
sonar	13.0	23.2	15.4	28.5	33.4	19.2	15.9	20.9	15.5	26.1	31.4	27.8	33.3
tao	4.6	3.6	5.1	7.0	7.4	5.6	4.3	6.4	4.7	7.0	6.3	8.4	7.2
thy2c3	2.8	3.1	2.9	6.6	11.3	4.9	4.4	7.0	5.4	7.2	7.8	8.6	10.5
vehicle2c1	6.6	13.3	8.7	12.1	14.9	10.4	7.7	10.8	7.3	13.1	13.7	13.0	14.0
vehicle2c2	24.7	25.6	24.2	25.2	27.6	24.9	24.8	25.2	23.6	26.8	27.4	26.5	27.8
vehicle2c3	26.1	27.8	26.9	28.8	28.2	27.8	26.7	27.3	25.9	27.4	29.3	29.0	27.9
vehicle2c4	4.0	11.6	6.3	11.4	14.0	7.7	5.1	7.7	5.5	11.9	11.4	12.3	13.5
wav2c1	16.8	18.7	17.3	18.7	19.0	18.0	17.0	17.7	17.0	4.8	18.6	4.8	4.9
wav2c2	19.8	21.2	20.6	21.6	21.6	20.9	19.7	21.3	19.9	21.5	21.6	21.6	21.9
wav2c3	16.4	17.5	16.2	17.6	18.2	16.9	16.6	16.9	16.5	17.5	18.4	17.8	18.7
wbcd	4.7	5.3	4.5	5.2	4.4	4.7	4.5	4.8	4.4	5.2	4.7	6.0	5.2
wisconsin	3.9	3.2	4.2	4.9	4.8	4.3	4.1	4.6	4.1	4.5	4.8	4.7	5.4
wpbc	26.3	34.1	29.0	27.9	28.0	29.0	27.5	29.5	28.3	30.1	26.6	28.0	28.9
Mitjana	14.9	16.9	15.5	18.0	18.8	16.5	15.2	16.8	15.3	17.5	18.4	17.7	18.4

Taula 7.13: Resultats per a l'aplicació de les estratègies desenvolupades a [2] sobre el conjunt de problemes de prova amb complexitat mitjana, segons la definició donada a l'apartat 4.7. El valor que es mostra és el percentatge d'error de classificació de l'algorisme i avaluat sobre el problema de prova j.

the Best Neighbours, *Part of the Elements from the Best Neighbours* i *an Opportunity for All the Neighbours*). Els valors 05 i 08 es refereixen al llindar en la recuperació, M_3 vol dir que s'han escollit elements de 3 veïns com a màxim, i N indica un procés de normalització a l'hora de determinar el pes de cada element de la memòria en la recuperació.

Algorisme	Estratègia
X_1	CBR
X_2	S_OBM
X_3	$S\_EBN\_M\_3$
X_4	$S\_PEBN\_M\_3$
X_5	S_PEBN
X_6	$S\_OAN\_05\_M\_3$
X_7	S_OAN_05
X_8	$S\_OAN\_08\_M\_3$
X_9	S_OAN_08
X_{10}	$S\_OAN\_05\_M\_3\_N$
X_{11}	$S\_OAN\_05\_N$
X_{12}	$S\_OAN\_08\_M\_3\_N$
X_{13}	$S\_OAN\_08\_N$

Taula 7.14: Explicació de l'estratègia a la qual correspon cada un dels algorismes assajats sobre els problemes de prova "tipus B", amb els resultats obtinguts mostrats a la taula 7.12.

En primer lloc, es realitza un anàlisi de Friedman sobre el conjunt de les dades, un cop determinats els rangs, amb el qual s'obtenen els resultats que es mostren a la taula 7.15. Dels valors dels estadístics obtinguts ($\chi^2_{F,cor}$ i F_I) s'observa clarament com és possible rebutjar la hipòtesi nul·la global: per tant, existeix com a mínim una comparació simple en què la diferència de comportament és significativa.

Un cop rebutjada aquesta hipòtesi, es planteja el conjunt de test a posteriori per estudiar quina o quines comparacions retornen diferències significatives i, per tant, han permès el rebuig d'H_0. Una primera opció, com s'ha dit, és calcular el valor de CD realitzant totes les possibles comparacions per parelles simples, i per tant utilitzant el valor que el test de Nemenyi dóna per CD:

$$CD_N = 3.33 \tag{7.29}$$

Això vol dir que qualsevol parella d'algorismes, els resultats dels quals es

	$M = 13,\ N = 30$
χ^2_F	176.06
C	0.99
$\chi^2_{F,cor}$	177.41
df	12
$\chi^2_{F,.05}$	21.03
Rebuig H_0	Si
F_I	27.76
df	12 x 348
$F_{I,.05}$	1.78
Rebuig H_0	Si

Taula 7.15: Aplicació del test de Friedman sobre els resultats de les estratègies de clusteritzacio exposades a [2], amb els problemes de prova de complexitat mitjana. Els valors de l'estadístic obtinguts sí permeten rebutjar la hipòtesi nul·la global.

diferenciïn de més de 3.33 en el valor del rang, poden ser considerats com amb comportament diferent. Aquest valor marca tot un conjunt de regions que determinen quins algorismes tenen comportament similars a quins. Una bona manera de veure-ho gràficament, és marcant sobre l'eix de valors de rang l'obtingut per cada algorisme, i connectant aquells amb comportaments no significativament diferents (seguint l'esquema gràfic proposat a l'apartat 5.2).

En aquest cas, per exemple, es pot veure com hi ha un conjunt de quatre algorismes (S_OAN_05 (X_7), S_OAN_08 (X_9), $S\_EBN\_M\_3$ (X_3) i $S\_OAN\_05\_M\_3$ (X_6)) que no mostren un comportament significativament diferent al del CBR. Per la resta d'algorismes, les estratègies proposades empitjoren el resultat obtingut de manera significativa.

A la figura 7.3 es mostra una possible representació gràfica d'aquest resultat: sobre un eix en què apareixen els valors dels rangs mitjans de cada algorisme (X_i), en sentit creixent, s'uneixen aquells valors que difereixen menys que el valor de CD_N calculat. Els algorismes (en aquest cas de A_1 a A_{13}) units són aquells que no tenen un comportament significativament diferent entre ells.

La segona opció d'anàlisi és a través d'una comparació amb l'algorisme de control, en aquest cas el CBR (amb valor mitjà de rang, X_1 a la figura, que com es veu no és pas el millor resultat obtingut): l'objectiu és determinar quins algorismes no es comporten de manera significativament diferent a

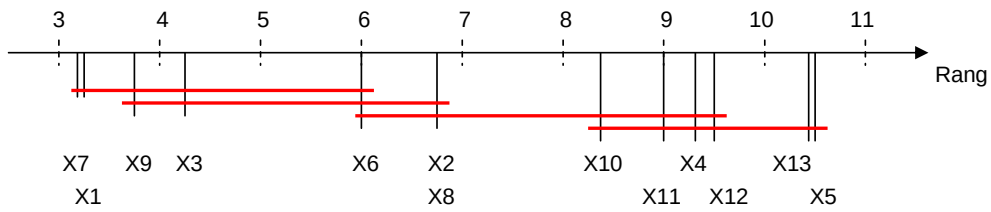

Figura 7.3: Representació gràfica dels rangs mitjans de cada algorisme. Les línies vermelles uneixen aquells algorismes que no mostren un comportament significativament diferent entre ells.

aquest. D'acord amb el plantejament original del treball citat, en què l'objectiu és reduir el número d'operacions en la fase de recuperació sense perdre precisió, aconseguir determinar en quins casos es complexi aquesta darrera condició és essencial. Amb un test de Bonferroni-Dunn, s'obté un valor de CD lleugerament inferior, com era d'esperar:

$$CD_{B|D} = 2.87 \tag{7.30}$$

d'ón es conclou el mateix que amb el test de Nemenyi, tot i que amb dubtes sobre l'algorisme $S\_OAN\_05\_M\_3$ (X_6): la diferència entre el seu rang i el rang del CBR és exactament el valor de CD i, per tant, estem al límit de la significança. La figura 7.4 mostra el gràfic equivalent al descrit anteriorment, ara per al valor que correspon a CD_{BD}.

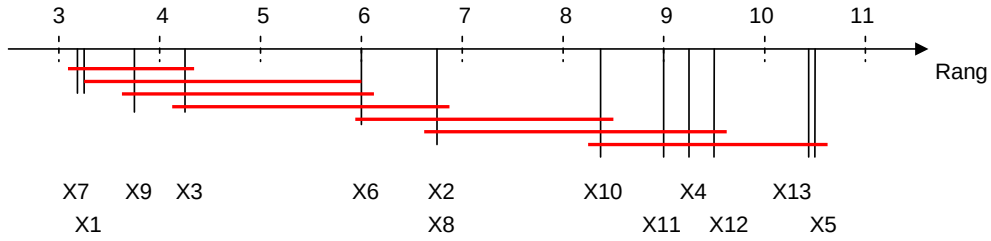

Figura 7.4: Representació gràfica dels rangs mitjans de cada algorisme. Les línies vermelles uneixen aquells algorismes que no mostren un comportament significativament diferent entre ells. Tal i com s'ha exposat al text, la diferència entre X_1 i X_2 coorespon exactament al valor de CD_{BD} calculat.

El dubte sobre l'algorisme $S\_OAN\_05\_M\_3$ (X_6) es resolt aplicant el mètode de Holm, com es descriu a la taula 7.16: el valor de p corresponent

a aquest algorisme (p_9) és clarament inferior a α_9, i per tant es pot concloure que hi ha evidències suficients per rebutjar l'opció que prediria igual comportament que el CBR. Aquest resultat constata el que ja s'havia dit anteriorment: per un cas de comparació respecte un algorisme de control, el mètode de Holm té una capacitat superior per discriminar diferències significatives, respecte al test de Bonferroni-Dunn.

i	**Alg. comp.**	$R_i - R_0$	z_i	p_i	$\frac{\alpha}{(M-i)}$
1	S_PEBN	7.30	7.260	< 0.001	0.0042
2	$S\_OAN\_08\_N$	7.25	7.210	< 0.001	0.0045
3	$S\_OAN\_08\_M\_3\_N$	6.20	6.166	< 0.001	0.0050
4	$S\_PEBN\_M\_3$	6.07	6.033	< 0.001	0.0056
5	$S\_OAN\_05\_N$	5.88	5.851	< 0.001	0.0063
6	$S\_OAN\_05\_M\_3\_N$	5.13	5.105	< 0.001	0.0071
7	$S\_OAN\_08\_M\_3$	3.53	3.514	< 0.001	0.0083
8	S_OBM	3.53	3.514	< 0.001	0.0100
9	$S\_OAN\_05\_M\_3$	2.87	2.851	0.0069	0.0125
10	$S\_EBN\_M\_3$	1.07	1.061	0.2273	0.0167
11	S_OAN_08	0.58	0.580	0.3372	0.0250
12	S_OAN_05	-0.02	-0.016	0.3989	0.0500

Taula 7.16: Aplicació del mètode de Holm sobre els resultats de les estratègies de clusterització exposades a [2], sobre els problemes de prova de complexitat mitjana. Els valors obtinguts, a partir de la comparació amb el CBR (de rang R_0), permeten rebutjar la hipòtesi d'igualtat de comportament pels 9 algorismes amb valor de p menor. Com de costum, el valor de confiança utilitzat és $\alpha = 0.05$.

7.5 Protocol d'aplicació de test per comparacions múltiples

El conjunt de tècniques exposades fins aquest moment, junt amb les propostes d'utilització que s'han anat presentant, permeten l'elaboració d'un protocol global d'actuació que sigui vàlid per qualsevol problema d'assaig d'M algorismes sobre N problemes de prova.

En primer lloc, cal discutir el compliment de les diferents condicions necessàries per aplicar l'anàlisi de variàncies, d'acord amb el ja proposat a l'apartat 7.3.2. En cas que sigui possible aquest anàlisi, el test de la hipòtesi nul·la global presenta dues possibilitats: si no es pot rebutjar H_0, l'anàlisi

finalitza aquí, doncs senzillament cal concloure que no hi ha diferències significatives entre els comportament dels M algorismes.

En canvi, si H_0 es pot rebutjar (d'acord amb els valors de F i F_α), cal plantejar-se tot el conjunt de comparacions $A_i \sim A_j$ que es volen realitzar (per parelles o complexes, totes les possibles o només els contrastos, etc.) i sobre cada una d'aquestes comparacions realitzar els càlculs de CD_{LSD} i $CD_{B|D}$. A continuació, es comprova la relació entre aquests valors i el valor de $\overline{X_i} - \overline{X_j}$, actuant d'acord a com s'indica en l'esquema.

En el cas que es compleixi que $CD_{LSD} < \overline{X_i} - \overline{X_j} < CD_{B|D}$, l'anàlisi paramètric no ens permet dir res i, a partir de la conclusió prèvia del rebuig d'H_0, la solució és anar a un test a posteriori no-paramètric, intentant que aquest ens pugui resoldre la incògnita. Per la forma com aquest darrer està plantejat, sense opció de no-conclusió final, és possible que no es detecti cap diferència significativa en $A_i \sim A_j$, per la menor capacitat de discriminació que acostuma a mostrar un mètode no-paramètric. Ja s'ha dit anteriorment que no és descartable que, malgrat es rebutgi H_0, no es pugui trobar la comparació o comparacions que provoquen aquets rebuig.

Finalment, si no es compleixen les condicions d'aplicabilitat de l'anàlisi de variàncies, es proposa optar per un test de Friedman, seguint després el mateix esquema que per l'anterior: en cas que no es pugui rebutjar H_0, l'anàlisi finalitza sense trobar cap diferència significativa entre els M algorismes assajats. Pel contrari, si es rebutja H_0 cal plantejar les comparacions a realitzar i optar per dues alternatives en funció del tipus de comparacions: si es tracta d'una comparació planificada o la comparació sobre un control és millor optar per un test de Holm, mentre que si s'analitzen totes les comparacions possibles (no planificades), es recomana un test de Nemenyi.

Les virtuds d'aquest protocol, que es representa a l'esquema que es mostra a la figura 7.5, són múltiples: d'una banda, clarifica l'ús dels diferents test que habitualment s'utilitzen per a comparar el comportament d'un conjunt d'algorismes, deixant clares les condicions per aplicar uns o altres; d'altra banda, no deixa mai el problema sense resposta, optant per un plantejament considerablement conservador, especialment en l'anàlisi dels valors de la distància crítica (CD). És possible que algun cop es cometi un error "Tipus II" (que no es detecti com a diferent el comportament de dos algorismes que sí que ho són), però es minimitza considerablement la possibilitat de cometre un error de "Tipus I" (la consideració de diferents per a dos algorismes que tenen el mateix comportament).

Per contra, és evident que seguir aquest protocol implica una gran quantitat de càlculs, molt superior als que habitualment es fan per a l'anàlisi desl

resultats en aquests problemes. Un dels objectius del treball és haver demostrat la necessitat de realitzar tots aquests càlculs, si el que es desitja és una conclusió sobre el comportament dels algorismes totalment fiable, a un determinat nivell de significança α.

7.6 Resum

L'objectiu principal d'aquest capítol era determinar una metodologia per a la comparació múltiple de resultats, és a dir, per a l'estudi del comportament d'un número $M > 2$ d'algorismes, a partir de les dades obtingudes de l'assaig d'aquests sobre una col·lecció d'N problemes de prova. Per a poder-ho fer, en primer lloc s'ha discutit la manera com establir una hipòtesi nul·la global, i com determinar i controlar el nivell de significança.

Un cop determinats aquests aspectes imprescindibles per a procedir a la comparació, s'han desenvolupat diferents qüestions fins arribar a un protocol final que determina com procedir en el cas d'una comparació de més de dos algorismes. En primer lloc cal discutir les condicions que estableixen el domini d'ús de l'anàlisi de variàncies tal i com s'ha exposat en el text. A partir d'aquí, si els resultats obtinguts són suficients com per permetre'n el seu ús es discuteix sobre la hipòtesi nul·la global, i si és rebutjada es procedeix a l'aplicació del test a posteriori necessari per a determinar el rang de valors de la distància crítica. L'anàlisi d'aquests valors, en comparació del resultats obtinguts, permet determinar quins algorismes mostren diferències significatives.

Si les condicions del domini d'ús abans esmentat no es compleixen, el protocol força a utilitzar un test no paramètric per discutir H_0 (en cas que aquesta sigui rebutjada), aplicable depenent de quin sigui l'objecte d'estudi. Aquests darrers passos donen també resposta a aquells casos en què, tot i poder aplicar els mètodes paramètrics, els resultats llavors obtinguts no permetin trobar una conclusió clara.

D'aquesta manera, es compleix el principal objectiu plantejat per aquest capítol: obtenir un protocol d'actuació per al cas de la comparació múltiple de resultats, que aporti sempre una resposta sobre les hipòtesis que es plantegin, i que ho faci controlant correctament el nivell de significança estadística (amb un tarannà conservador alhora d'extreure'n conclusions), i sense descartar l'aplicació de l'anàlisi de variàncies, amagant-se darrera de suposades dificultats en el càlcul de les condicions per a determinar la seva aplicabilitat.

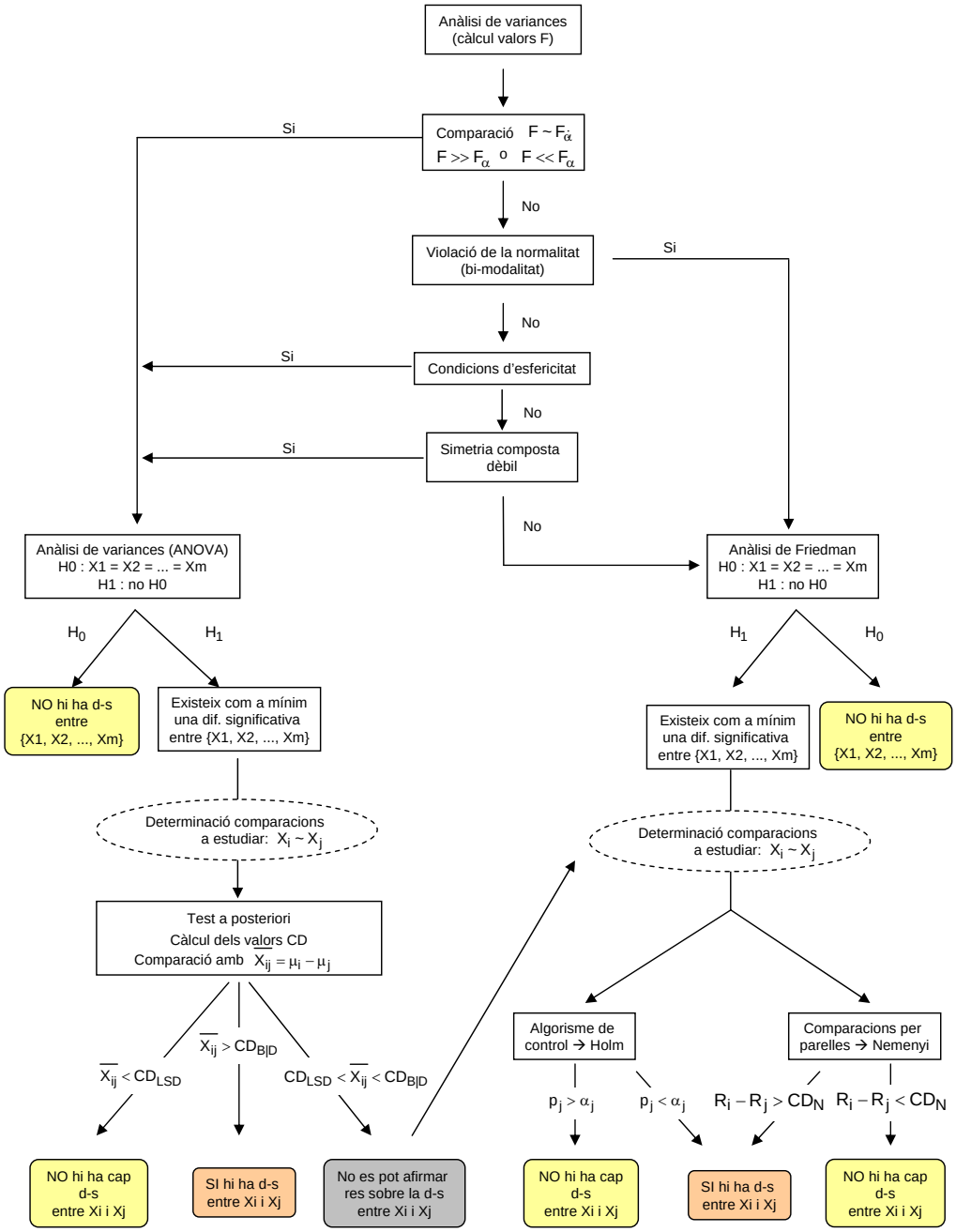

Figura 7.5: Proposta de protocol per a la correcta aplicació dels test per comparacions múltiples, considerant M algorismes ($M > 2$) sobre N problemes de prova. El color diferencia els casos en què es pot rebutjar la hipòtesi nul·la (i, per tant, hi ha diferència significativa, d-s) d'aquells en què no.

Part IV

Avaluació de les metodologies i conclusions

Capítol 8

Avaluació de les metodologies

"Tests performed by different scientist with the same pair of algorithms, the same data sets and the same hypothesis test may present different results."
Remco R. Bouckaert

En els capítols precedents, s'han estudiat diferents metodologies per avaluar el comportament d'M algorismes, des d'un punt de vista comparatiu i a partir d'una mesura de bondat determinada. En aquest apartat es canvia l'orientació de la qüestió, i s'estudien eines per avaluar la bondat de les pròpies metodologies, que en el cas estudiat (comparacions simples) són diversos test d'inferència estadística (t-test, Wilkoxon, Binomial).

D'acord amb el comentat fins ara, l'elecció d'una o altra metodologia es basa essencialment en dos criteris: d'una banda, el compliment de les condicions d'ús establertes per cada test; d'altra banda, el fet que si es compleixen les condicions que en garanteixen una conclusió fiable, un test paramètric sempre serà preferible a un no-paramètric, doncs es tindrà en consideració major informació, provinent dels resultats obtinguts sobre la col·lecció de problemes de prova que correspongui. Aquestes dues qüestions determinen quin test utilitzar, com es resumeix també a la figura 6.2, en el cas de les comparacions simples.

La pregunta fonamental en aquest punt és: per quin motiu una metodologia serà millor o pitjor que una altra? Per una part, és evident que cal que les conclusions que aporti siguin fiables (i això ho garanteix el seu domini d'ús). Per altra part, es desitja un test amb una gran capacitat per determinar quan hi hagi una diferència significativa entre els dos algorismes, però que

a la vegada això no impliqui un augment de la probabilitat de rebutjar una hipòtesi nul·la quan és certa (relació entre l'error "Tipus I" i l'error "Tipus II", discutida al capítol 5).

Doncs bé, l'objectiu del present capítol és discutir més profundament aquestes qüestions, i determinar numèricament uns indicadors que permetin avaluar la segona qüestió, i discutir si hi ha o no relació entre ambdues: és a dir, si el compliment de les condicions del domini d'ús d'un test determina la seva capacitat per determinar l'existència d'una diferència significativa (com es veurà al següent apartat). Per simplicitat, els resultats s'obtindran per casos en què es comparen dos algorismes entre sí (comparacions simples), però tots els raonaments serien extrapolables a comparacions múltiples.

A banda del comentat, hi ha un altre factor que pot determinar la bondat d'una metodologia per avaluar els resultats obtinguts. Cal tenir en compte que un mateix test d'inferència estadística, aplicat sobre els resultats obtinguts per dos algorismes sobre una mateixa col·lecció de problemes de prova, pot concloure diferent si s'aplica més d'una vegada. El test en qüestió serà tan més interessant per la seva aplicació com més coherents siguin les conclusions que aporta: direm que es desitja un test amb una elevada replicabilitat, i aquest altre factor serà l'estudiar a l'apartat 8.2.

8.1 Potència d'un test en comparacions simples

La primera qüestió que s'estudiarà és la capacitat d'una determinada metodologia d'inferència estadística per trobar diferències significatives entre dos algorismes, quan aquestes existeixin. En aquesta línia, es defineix la potència d'un test d'inferència estadística com la capacitat d'aquest per determinar l'existència d'una diferència significativa, en cas que aquesta existeixi. Formalment parlant, la potència és igual a la probabilitat que el test rebutgi correctament la hipòtesi nul·la.

8.1.1 Definició i procediment per al càlcul de la potència

A l'hora de definir-la, la potència[1] d'un test és complementària de la probabilitat de cometre un error de "tipus II" (β), definit com la probabilitat de no

[1] Malgrat no sigui un terme molt indicatiu, es manté la traducció directe de l'habitual *power*.

rebutjar la hipòtesi nul·la quan realment existeix una diferència significativa:

$$pot = 1 - \beta \tag{8.1}$$

En l'anàlisi sobre una hipòtesi nul·la (H_0), la probabilitat de cometre un error "Tipus I" ve determinat pel nivell de significança α, de tal manera que un valor d'α molt proper a 0 implica una probabilitat molt petita de preveure una diferència significativa quan aquesta no hi és. Tot i això, si α es redueix a valors molt petits, aleshores augmenta de manera significativa β, amb la qual cosa es redueix la potència del test utilitzat per a l'anàlisi de H_0. Per tant, α i la potència d'un test estan estretament relacionats per a cada problema particular, essent en certa mesura proporcionals: per això, a l'hora d'avaluar la potència entre un conjunt de metodologies o test, l'anàlisi es farà a partir de fixar un valor d'α, igual per tots ells, i comparar-ne la corresponent potència.

Per fer aquesta anàlisi, i en el cas en què es tinguin 2 algorismes assajats sobre un conjunt de N problemes de prova, el més habitual és forçar la diferència en la bondat d'ambdós algorismes a partir d'una certa parametrització, i mesurar com augmenta la potència de cada test en relació a aquesta diferència. La comparació d'aquesta evolució per cada test aporta una mesura de, com a mínim, la potència relativa de cada un d'ells respecte els altres.

Una possibilitat per efectuar aquest càlcul és la tècnica definida per Dietterich ([13]), on s'altera artificialment el comportament d'un dels dos algorismes que es comparen per induir-hi errors, i d'aquesta manera mesurar la capacitat dels test per detectar aquesta diferència significativa. L'alteració d'un dels dos algorismes es pot regular per un paràmetre, que permet tenir un indicador de quina és la diferència *real* entre ambdós, i estudiar així l'augment de la potència del test corresponent en funció de com evoluciona amb aquest paràmetre introduït.

Aquesta opció té diversos problemes: d'una banda, requereix assajar l'algorisme sobre tots els problemes de prova per cada valor diferent del paràmetre que indica quin error s'ha generat i, per tant, no és òptim a nivell computacional; l'altra és que els errors són provocats tot alterant el propi funcionament de l'algorisme i, per tant, no s'està estudiant estrictament el mateix algorisme en cada assaig. Per superar aquestes dificultats, la millor opció és actuar de manera similar com ho fa Demsar ([9]): els algorismes s'assagen sobre una col·lecció amb $N' < N$ problemes de prova, de tal manera que les modificacions es fan sobre l'elecció d'aquests N' problemes, i es mantenen inalterats els algorismes estudiats, canviant la col·lecció sobre la qual s'avaluen.

Només cal, doncs, parametritzar l'elecció dels problemes en funció de la diferència de comportament entre els dos algorismes comparats: es defineix, en primer lloc, Δ_j com la diferència en el resultat obtingut de l'assaig dels dos algorismes sobre el problema de prova j. A partir d'aquí, s'avalua el resultat obtingut per una nova col·lecció de N' problemes, en què la probabilitat d'incloure el problema de prova j en aquesta col·lecció és proporcional al resultat del terme

$$\frac{1}{1+e^{-k\Delta_j}} \tag{8.2}$$

El paràmetre k que apareix en aquesta expressió determina quin és l'efecte de la diferència dels resultats en l'elecció d'un problema de prova o altra. Així, i seguint les formes que es mostren en la figura 8.1, un valor de k proper a 0 implica una probabilitat similar per tots es problemes de ser escollits per formar part de la col·lecció en què s'assagen els algorismes, doncs el terme 8.2 pren un valor similar per qualsevol Δ_j. En aquest cas, la diferència entre ambdós algorismes no serà significativament diferent a la trobada en l'assaig sobre tot el conjunt dels problemes, i ho fa sense alterar els algorismes en sí mateixos.

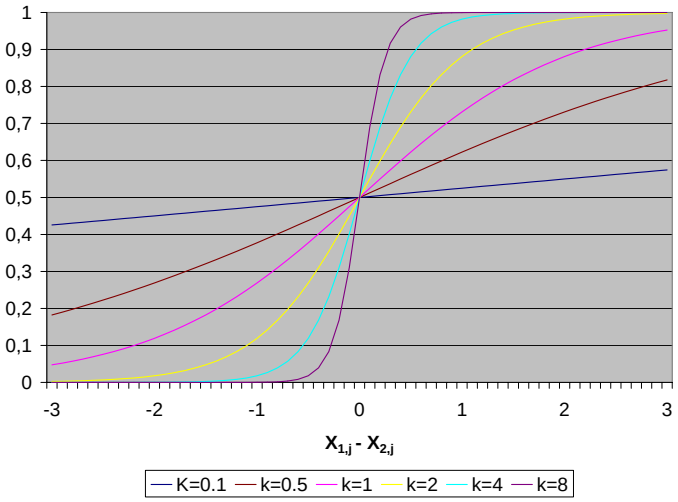

Figura 8.1: Representació del factor de probabilitat d'elecció definit a l'equació 8.2, per diferents valors de k. Els valor en l'eix d'abscisses representen la diferència entre els resultat obtingut pels dos algorismes sobre un problema de prova determinat, $X_{1,j} - X_{2,j} = \Delta_j$. Es veu com per valors de k propers a 0 el valor obtingut varia lentament amb Δ_j, mentre que per valors elevats de k la probabilitat d'elecció varia molt ràpidament quan ho fa Δ_j.

En canvi, per valors elevats de k, la forma de la funció proposada implica que aquells problemes en què Δ_j és més elevat tinguin més probabilitats de ser escollits pera l'assaig: per tant, valor elevats del paràmetre k implica assajar els algorismes sobre una col·lecció de problemes en què les diferències de comportament seran elevades. Per això es pot afirmar que k parametritza la diferència de comportament entre els algorismes.

A banda dels elements comentats, cal tenir en compte també que aquest procediment supera les dues dificultats detectades en la proposta de Dietterich: no requereix l'assaig de nous algorismes i, per tant, el temps de càlcul necessari és molt menor, i a més els algorismes no es veuen en cap moment alterats. Es pot estar segur que la comparació és entre els algorismes desitjats.

A partir d'aquí, el procediment és com segueix:

1. Es defineix el valor de la ignificància α per a tot el problema. Habitualment, $\alpha = 0.05$.

2. Per cada valor de k, començant per 0 i en sentit creixent, es generen L col·leccions de $N' < N$ problemes de prova.

3. Aquests problemes de prova s'escullen d'acord amb el fet que la probabilitat que un problema j sigui membre d'aquesta col·lecció és proporcional a l'expressió 8.2.

4. Per cada una d'aquestes L col·leccions, es mesura el valor de p: la probabilitat que la diferència obtinguda en la col·lecció de problemes de prova generada sigui deguda a l'atzar i que, per tant, els dos algorismes tinguin el mateix comportament.

5. Finalment, s'avalua el valor mig dels L valors de p obtinguts, $\bar{p}$, i s'estudia l'evolució d'aquesta magnitud respecte el valor del paràmetre que caracteritza la diferència entre els dos algorismes, k.

Com menor sigui el valor de $\bar{p}$ obtingut, determinat un nivell de significança α, major serà la probabilitat de rebutjar la hipòtesi nul·la si existeix una diferència significativa entre els dos algorismes. Això vol dir que valors petits de $\bar{p}$ impliquen una elevada potència del test d'inferència utilitzat per a comparar els dos algorismes. El procediment permet obtenir l'evolució de $\bar{p}$ respecte k per cada test i, per tant, comparar així la potència de cada un d'ells.

Sobre el procediment definit, cal comentar en especial el primer dels passos: de quina manera es procedeix a l'elecció del conjunt de N' problemes de prova, si es coneix la probabilitat d'un d'ells de ser escollit, proporcional a l'expressat al terme 8.2. És qüestió de generar L tries d'N' elements d'un conjunt d'N elements, sense que tots ells siguin equiprobables. Un cop normalitzats tots els valors obtinguts, per tal que la probabilitat total sumi 1, l'elecció es realitza a partir del mètode de la rejecció, que es basa en la generació d'una tria a partir d'una altra funció de probabilitat coneguda, i la posterior modificació de la tria a partir de la relació entre ambdues funcions de probabilitat ([135], [136]). La metodologia, relacionada amb les tècniques de Montecarlo, no és senzilla d'implementar, i afegeix una nova dificultat al càlcul que s'està realitzant.

8.1.2 Comprovació de la relació entre la potència i el domini d'ús

La introducció de la potència com a indicador de la capacitat d'un test de determinar diferències significatives, quan aquestes existeixen, busca donar una eina complementària per a la discussió del test d'inferència a aplicar en un problema determinat. Com s'ha comentat anteriorment, fins ara el domini d'ús i la quantitat d'informació utilitzada determina el test a aplicar. La hipòtesi que es presenta és que les conclusions que s'obtenen del càlcul de la potència dels test són coherents amb les condicions del domini d'ús del mateix.

La veracitat d'aquesta afirmació es comprova en el següent exemple. En primer lloc, es consideren els $N = 18$ problemes de prova de domini mèdics que s'utilitzen per a l'assaig de les variacions del SOMCBR presentades a [2], i que ja s'han mostrat a la taula 6.6. A partir d'aquests problemes s'han construit, per cada valor de $k \in (0, 10)$, $L = 1000$ col·leccions de $N' = 8$ problemes de prova, per les quals s'han avaluat els valors de la probabilitat p i, a partir dels L resultats obtinguts, el valor de $\bar{p}$.

Aquest procediment s'ha realitzat per la comparació per parelles dels algorismes $OAN\_08\,(X_1)$, $OAN\_05\_MAX\_NORM\,(X_2)$ i OAN_05_NORM (X_3). En totes tres comparacions (X_1 respecte X_2, X_1 respecte X_3 i X_2 respecte X_3), s'han avaluat els valors de $\bar{p}$ per al t-test, el test de signes de Wilkoxon i el test binomial, amb l'objectiu de comparar la potència dels tres principals test presentats en el capítol 6.

Com també es vol comparar aquesta magnitud amb el compliment de

les condicions del domini d'ús de cada un dels test, convé recordar que a la taula 6.7 s'hi mostren els valors dels paràmetres que determinen aquest compliment (caràcter normal de les dades, homogeneïtat de variàncies, valor de l'estadístic obtingut en relació amb el crític). En aquells apartats s'ha conclòs que el t-test està dins el seu domini d'ús per la comparació entre X_1 i X_2, però no per altres dues (X_1 i X_3, X_2 i X_3).

Pel que fa al valor de la potència mitjana, $\bar{p}$, els resultats es mostren a la figura 8.2, i han de ser analitzats separadament per cada comparació, doncs el compliment de les condicions d'ús dels test és diferent en cada cas.

Pel que fa a la comparació entre X_1 i X_2 (gràfica (a) de la figura), cal considerar en primer lloc que no es compleixen totes les condicions necessàries per a poder aplicar el t-test, doncs no es pot assegurar el caràcter normal de la distribució de Δ_j, tal i com ja s'ha discutit a partir de la taula 6.7. Tot i això, el compliment de la homogeneïtat de variàncies i el fet que el valor de $\bar{p}$ calculat pel t-test compleixi $p << \alpha$ permet considerar que s'està en un cas en què les conclusions del t-test seran fiables: s'està dins el domini d'ús d'aquest test.

De manera coherent amb això, a la gràfica s'observa com a partir d'un cert valor de k (per tant, a partir d'una certa diferència important entre els algorismes comparats) el valor de $\bar{p}$ es redueix acostant-se a 0 ràpidament, i els valors obtinguts pel t-test sempre són menors que els altres. De fet, pels valors de k en què la forma de l'expressió 8.2 indueix diferències representatives ($k > 1$), es compleix sempre la següent relació:

$$\bar{p}_{t-test} < \bar{p}_{Wilk} < \bar{p}_{Bin} \tag{8.3}$$

Aquesta relació és perfectament coherent amb la hipòtesi que s'havia plantejat: en aquells casos en què sigui possible aplicar el t-test, la seva potència serà major que els altres test no-paramètrics, doncs té en compte molta més informació en la determinació del corresponent estadístic. És el mateix motiu pel qual el test de Wilcoxon té una potència major que el binomial. Dit d'una altra manera, el que s'observa és el comportament esperat per aquells casos en què el t-test està dins el seu domini d'ús.

En els altres dos casos, en canvi, ja no es mostra tan clarament la diferència de resultats per $\bar{p}$ en els tres test assajats. De fet, per valors grans de k l'evolució de $\bar{p}_{t-test}$, $\bar{p}_{Wilk}$ i $\bar{p}_{Bin}$ és molt similar, i fins i tot en l'extrem el t-test es comporta pitjor que els altres dos. Això és perquè en aquests dos casos (X_1 respecte X_3, i X_2 respecte X_3), el t-test ja no està dins el seu domini d'ús (veure de nou la taula 6.7), i per tant és preferible la utilització

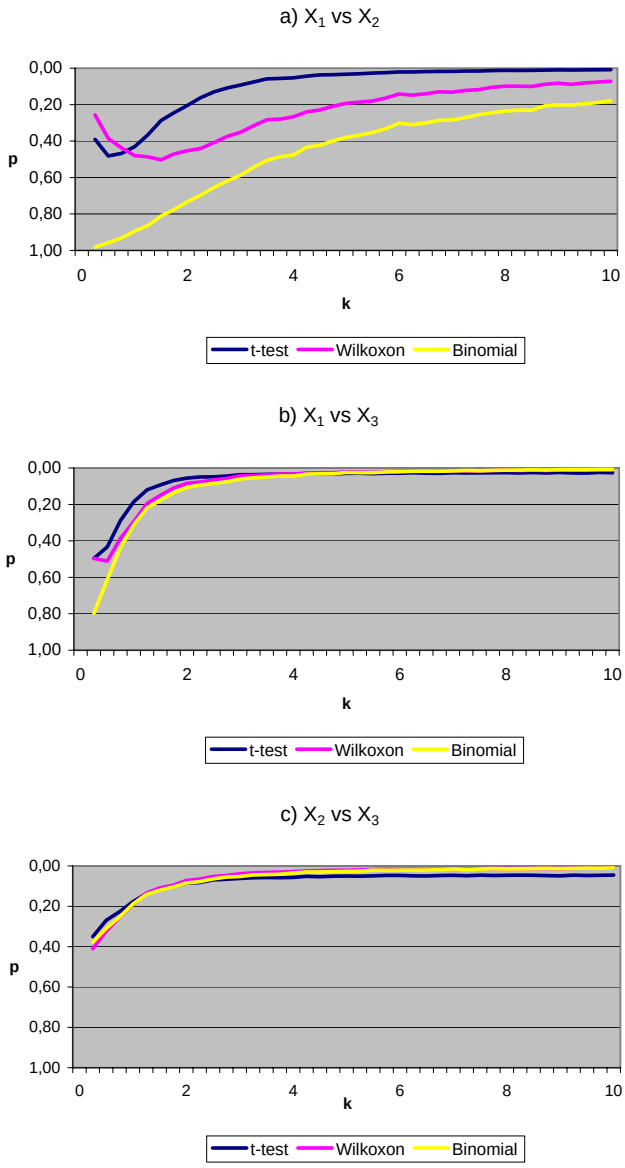

Figura 8.2: Evolució del valor mig de la probabilitat que la diferència obtinguda en la col·lecció de problemes de prova generada sigui deguda a efectes aleatoris i que, per tant, els dos algorismes tinguin el mateix comportament ($\bar{p}$), respecte el paràmetre que dóna una mesura de la diferència de comportament entre els algorismes (k). Els càlculs estan fets amb $L = 1000$ i $N' = 8$, per a les comparacions de X_1 amb X_2 (figura a), X_1 amb X_3 (figura b), i X_2 amb X_3 (figura c), sobre els $N = 18$ problemes de prova presentats a la taula 6.6.

d'un test no-paramètric, com s'ha exposat abastament a l'apartat 6.3. També en aquests casos, per tant, es fa palesa la relació entre el compliment de les condicions d'ús que permet confiar en la conclusió donada per un test, i el seu poder mesurat respecte l'augment de la diferència entre els algorismes.

Una altra mesura que també dóna una idea sobre la potència d'un test és la que s'expressa a continuació: si tenim en compte que la potència és la probabilitat de rebutjar correctament la hipòtesi nul·la (H_0), i que per valors de $k >> 0$ s'està induint, amb tota seguretat, una diferencia significativa sobre el resultat dels dos algorismes que es comparen, el número de vegades que es rebutja H_0 d'entre les L col·leccions de problemes pot ser considerat també un indicador d'aquesta potència.

A la figura 8.3 es mostren els resultats obtinguts, on H és el número de vegades que es rebutja H_0 (i, per tant, $H \leq L$, amb $L = 1000$ en aquest cas). Els resultats obtinguts són perfectament coherents amb l'evolució de $\bar{p}$ respecte k, i mostren encara amb més claredat la limitació del t-test per aquells casos en què no es troba dins el seu domini d'ús, com és el cas de la comparació de X_1 amb X_3 (gràfica (b) de la figura) i de X_2 amb X_3 (gràfica (c) de la figura). En canvi, si és dins la seva zona d'ús com a la la comparació de X_1 amb X_2 (gràfica (a) de la figura), el t-test és clarament la millor opció per a l'anàlisi de la comparació simple.

8.2 La replicabilitat d'un test en comparacions simples

Recentment, alguns autors ([41]) han posat en dubte que l'error de "Tipus I" i l'error de "Tipus II" (o la potència, definida com a l'apartat anterior) siguin mesures suficients per avaluar la idoneïtat d'un test, a l'hora de comparar la bondat de dos algorismes a partir de les mesures obtingudes en l'assaig sobre N problemes de prova. La proposta que fan tendeix a definir noves magnituds que aportin informació sobre la replicabilitat del test, definida abastament al següent apartat.

En l'apartat anterior s'ha mostrat com existeix una relació directe entre la potència del test i el compliment de les condicions que determinen el seu domini d'ús. De fet, es planteja la potència com una conseqüència de les restriccions per a la seva aplicació. En aquest apartat s'estudiarà, de manera similar, la hipòtesi segons la qual existeix una forta relació entre la replicabilitat i aquestes restriccions: és a dir, si la replicabilitat aporta realment nova informació sobre la idoneïtat d'un test o si, per contra, és simplement una nova conseqüència de les condicions que determinen el seu

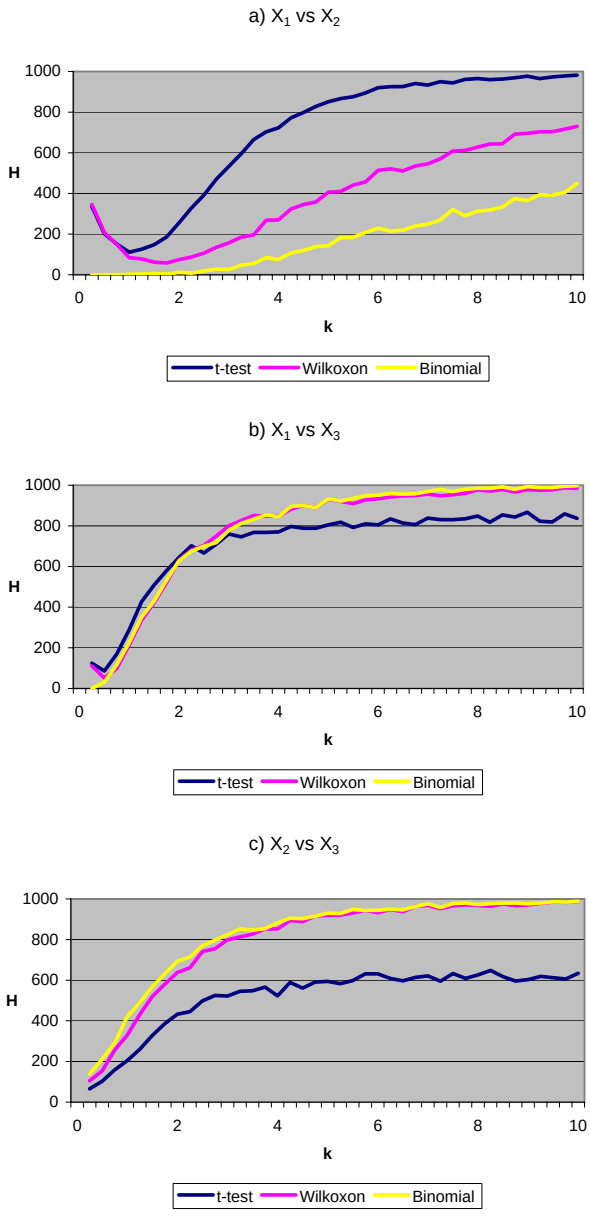

Figura 8.3: Evolució del número de casos en què es rebutja la hipòtesi nul·la d'entre les L col·leccions generades, amb $L = 1000$, respecte el paràmetre que dóna una mesura de la diferència de comportament entre els algorismes (k). Els càlculs estan fets amb $N' = 8$ sobre els $N = 18$ problemes de prova presentats a la taula 6.6, per a les comparacions de X_1 amb X_2 (figura a), X_1 amb X_3 (figura b), i X_2 amb X_3 (figura c).

domini d'ús.

8.2.1 Definició i procediment per al càlcul de la replicabilitat

En primer lloc, cal posar de manifest que les dades obtingudes de l'assaig d'un algorisme sobre una col·lecció de problemes de prova no és totalment determinista, doncs habitualment depèn de diversos factors aleatoris: des del propi mètode de partició del problema per dur a terme les fases d'entrenament i test, fins la possibilitat de llavors en el propi algorisme, hi ha diversos factors que trenquen una primera visió del problema més pròpia del principi de causalitat. És a dir, amb el mateix problema de prova, els mateixos algorismes A_1 i A_2 a comparar i el mateix test d'inferència estadística, dos experiments diferents podrien portar a conclusions oposades sobre una determinada hipòtesi nul·la.

Tenint en compte aquesta realitat, és lògic pensar que es prefereix aquell test que, per similar potència, provoqui un mínim en la probabilitat de tenir aquesta varietat de conclusions. Dit en altres paraules, és important analitzar la probabilitat que un test repeteixi la conclusió si s'aplica repetidament en la comparació de dos algorismes sobre una col·lecció de problemes de prova determinada.

Seguint aquest esquema, Bouckaert ([41]) defineix dues mesures relacionades amb el que es persegueix: la consistència i la replicabilitat. D'aquestes, la segona és molt més determinant i aporta més informació que la primera, i és per tant la que s'utilitza en aquest treball per estudiar la seva relació amb el compliment de les condicions que determinen el domini d'ús dels test. Per a comprovar la hipòtesi proposada, s'utilitza la definició de replicabilitat per la qual és igual a la probabilitat que dos experiments amb la mateixa parella d'algorismes arribi a la mateixa conclusió, és a dir, que dos experiments acceptin o rebutgin la hipòtesi nul·la.

Si es realitzen L assajos dels dos algorismes, dels quals en H casos la hipòtesi nul·la és rebutjada (i, per tant, acceptada en $L-H$ casos), la replicabilitat es pot calcular fàcilment segons la següent expressió:

$$R = \frac{H(H-1) + (L-H)(L-H-1)}{L(L-1)} \tag{8.4}$$

amb valors que van des de 0.5 fins a 1. Si el resultat obtingut és proper a 0.5 i, per tant, $H \simeq L - H$, el test conclou amb pràcticament la mateixa

freqüència l'acceptació o el rebuig d'H_0, amb la qual cosa queda desacreditat per a ser utilitzat en la comparació simple que es duu a terme. En canvi, si el valor de R s'acosta a 1 (és a dir, que el test mostra una alta replicabilitat) el test es manté molt coherent en les seves conclusions. Òbviament, seran preferibles sempre valors de R propers a 1.

8.2.2 Comprovació de la relació entre la replicabilitat i el domini d'ús

L'evolució d'aquest paràmetre respecte el valor de k (que s'ha definit en l'apartat anterior com una parametrització de la diferència entre els algorismes que es comparen) es mostra a la figura 8.4 per als tres algorismes utilitzats també en l'apartat anterior. Es mantenen els $N = 18$ problemes presentats a la taula 6.6 i l'assaig dels algorismes (X_1, X_2 i X_3) sobre una col·lecció de $N' < N$ problemes de prova, un total de $L = 1000$ vegades per cada valor de k.

En el primer gràfic d'aquesta figura es mostra l'evolució per a la comparació de X_1 amb X_2 per als tres test analitzats (t-test, Wilkoxon i binomial). S'observa com, per valor de k elevats, el t-test s'acosta a valors màxims de replicabilitat, mentre que els altres dos es mouen en regions properes a $R = 0.5$. Això és coherent amb el compliment de les condicions que determinen que el t-test està dins el seu domini d'ús (veure l'apartat anterior), mentre que en el test binomial s'observa un comportament a l'altre extrem: la informació que s'utilitza és tan poca que, per valors elevats de k, les seves conclusions sobre la hipòtesi del problema són pràcticament aleatòries.

En canvi, i de la mateixa manera que ja passava pel càlcul de la potència, en les altres dues comparacions (X_1 amb X_3 i X_2 amb X_3) els comportaments observats pels 3 test són molt diferents, d'acord amb el fet que el t-test ja no es troba dins el seu domini d'ús. Mentre que el test de Wilkoxon i el binomial augmenten la seva replicabilitat amb k, acostant-se a 1, el t-test es situa en valors molt més baixos. En la comparació de X_2 amb X_3 apareix el seu comportament extrem, amb un R proper a 0.5.

De nou, la hipòtesi plantejada es confirma: l'estudi estricte de les condicions que permeten aplicar el t-test, segons com s'ha definit el seu domini d'ús, aporta un coneixement sobre la fiabilitat de les conclusions que també és aportat per magnituds com la potència o la replicabilitat del test.

En cas d'haver calculat abans la potència a partir de la mitjana dels valors de p sobre L experiments, la replicabilitat també es útil interpretar-la com a

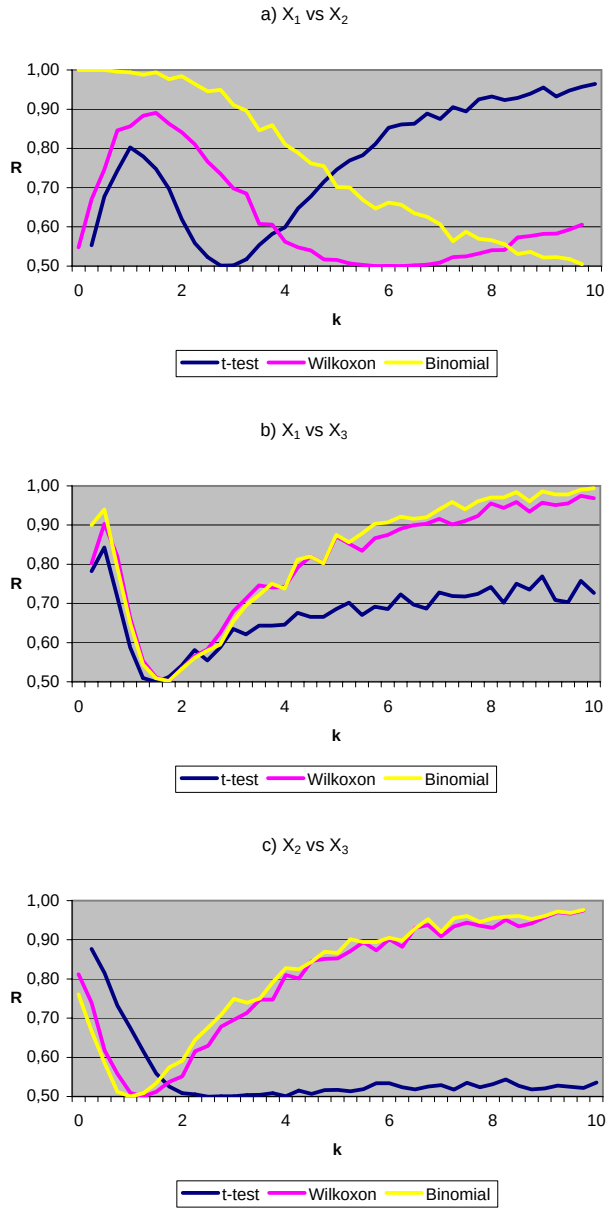

Figura 8.4: Evolució de la replicabilitat R respecte el paràmetre que dóna una mesura de la diferència de comportament entre els algorismes (k). Els càlculs estan fets sobre els $N = 18$ problemes de prova presentats a la taula 6.6, per a les comparacions de X_1 amb X_2 (figura a), X_1 amb X_3 (figura b), i X_2 amb X_3 (figura c).

inversa a la variació en els valors de p obtinguts en cada assaig: un test amb una alta replicabilitat proporcionarà valors de p sempre similars, mentre que un amb baixa replicabilitat trobarà valors de p amb elevada variació. Com la variança de p té un domini $[0, 0.25]$, Demsar ([9]) defineix una mesura alternativa de replicabilitat com

$$R(p) = 1 - 2Var[p] \tag{8.5}$$

Aquesta variable té un domini $[0.5, 1]$, i el propi Demsar demostra també que existeix una certa equivalència a la replicabilitat definida per Bouckaert, R.

Els resultats obtinguts per aquesta nova variable $R(p)$ es mostren a la figura 8.5, permetent conclusions en la línia del que proporciona el càlcul de R. Per la comparació entre X_1 i X_2, el t-tets està dins el seu domini d'ús, i és el que millor comportament mostra amb diferència: per valors elevats de k, la replicabilitat $R(p)$ tendeix ràpidament a 1, el seu valor màxim, mentre que el test de Wilkoxon i el binomial creixen també, però ho fan en menor mesura.

En canvi, pels altres dos casos (X_1 amb X_3, i X_2 i X_3) amb el t-test fora del seu domini d'ús, no hi ha pràcticament diferència i els tres tests tendeixen a una variança nul·la en el valor de p.

8.3 Resum

Tenint en compte les metodologies elaborades fins a l'inici d'aquest capítol, la tria d'un test per analitzar les diferències entre dos o més algorismes, a partir dels resultats obtinguts després d'aplicar-los sobre una col·lecció de problemes de prova, depenia de dos factors: el seu domini d'ús i la quantitat d'informació que consideraven.

El primer factor bé determinat, tant en el cas simple com en el múltiple, pels protocols d'ús desenvolupats als capítols 6 i 7, on es determina com estudiar el compliment de les restriccions per a l'ús d'un o altre test. L'altre factor admet una anàlisi molt més simple: quan sigui possible perquè així ho permeti el compliment de les citades restriccions, es preferirà sempre un test que tingui en compte quanta més informació millor per a extreure les seves conclusions. Aquest és el cas dels test paramètrics, que a diferència dels no paramètrics, treballen directament sobre els valors numèrics indicadors de la bondat de l'algorisme.

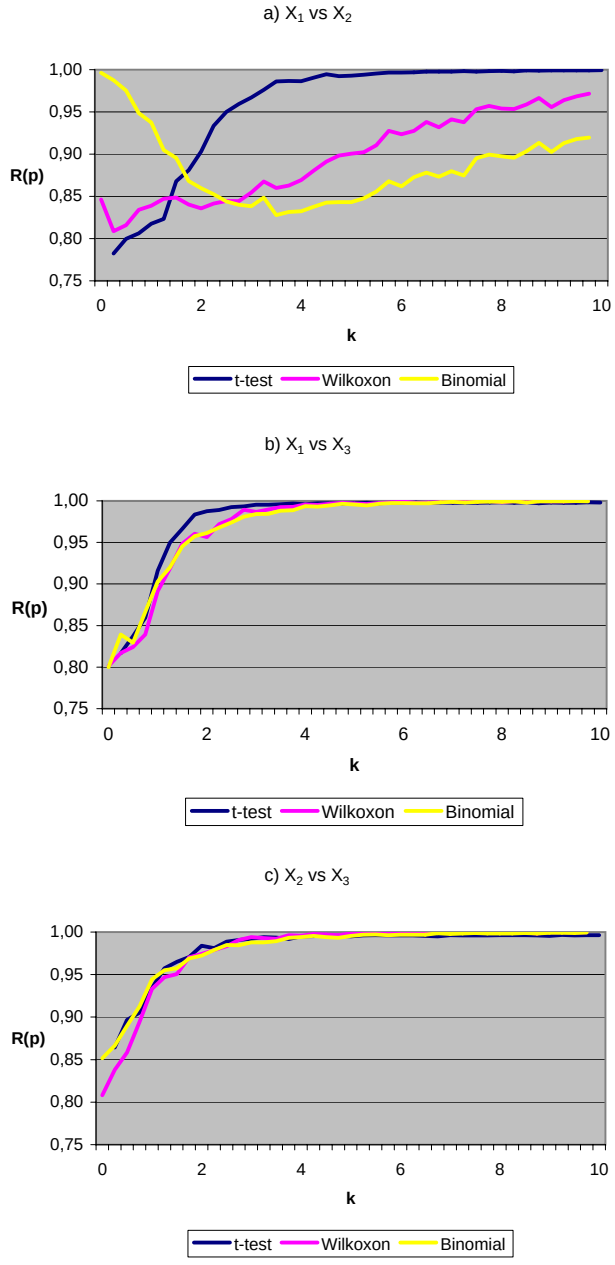

Figura 8.5: Evolució de la replicabilitat $R(p) = 1-2Var[p]$ respecte el paràmetre que dóna una mesura de la diferència de comportament entre els algorismes (k). Els càlculs estan fets sobre els $N = 18$ problemes de prova presentats a la taula 6.6, per a les comparacions de X_1 amb X_2 (figura a), X_1 amb X_3 (figura b), i X_2 amb X_3 (figura c).

En aquest capítol s'ha estudiat si per avaluar la bondat d'una metodologia de comparació cal tenir en compte algun altre factor, més enllà dels ja analitzats. En concret, s'han estudiat els indicadors que donen una idea de la potència d'un test (definida com la capacitat d'aquest per determinar l'existència d'una diferència significativa, en cas que aquesta existeixi), i els que aporten informació sobre la seva replicabilitat (definida com la probabilitat que un test repeteixi la conclusió si s'aplica repetidament en la comparació de dos algorismes, sobre una col·lecció de problemes de prova determinada).

L'estudi s'ha realitzat per test aplicables sobre comparacions simples i, per tant, s'han comparat el t-test, el test de Wilcoxon i el test binomial. La metodologia d'analisi però, seria exactament la mateixa si es comparessin els test aplicables sobre comparacions múltiples. A partir dels resultats obtinguts, que s'han representat a les figures 8.2 - 8.5, s'ha pogut concloure que la potència i la replicabilitat d'un test és conseqüència directe de les conclusions ja conegudes que determinen l'aplicació d'un o altre test: el compliment de les restriccions que conformen el domini d'ús, i la quantitat d'informació tinguda en compte pel propi test.

És a dir, en aquells casos en que el t-test era la millor opció segons aquests criteris, s'obté que també és la opció amb una major potència i una major replicabilitat. En canvi, quan no es compleixen les condicions per a la seva aplicació, els valors de potència replicabilitat obtinguts recomanen la utilització dels test no paramètrics, amb preferència pel test de Wilcoxon per sobre del test binomial.[2]

Dit d'una altre manera, el càlcul de les propietats estudiades en aquest capítol no és necessari per a determinar el test d'inferència a utilitzar, si es fan servir correctament els protocols definits als capítols 6 i 7. O de manera inversa, el càlcul d'aquests indicadors aporta el mateix resultat que si es segueixen els citats protocols.

[2]Coherentment amb el fet que el primer té en compte més informació que el segon per concloure sobre una hipòtesi nul·la.

Capítol 9

Aplicacions

Durant tot el treball s'han utilitzat resultats obtinguts prèviament per a exemplificar l'ús proposat de diferents tècniques d'anàlisi dels resultats. S'ha fet sense vocació de completitud, analitzant només la qüestió en concret que s'estava exposant, sense preocupar-se pel compliment de condicions prèvies o per la utilitat real de la seva aplicació en aquell cas.

Finalitzada l'exposició del conjunt de tècniques, i dels protocols proposats per a la seva correcta utilització, en aquest capítol es planteja la "resolució" complerta de cada problema. És a dir, donades les dades que expressen la bondat d'un conjunt d'M algorismes aplicats sobre N problemes de prova, es procedeix a la seva anàlisi complerta seguint els protocols proposats fins arribar a una conclusió final. Si és el cas, es compara la conclusió obtinguda amb la proposada en el seu moment, i es conclou sobre la validesa de la metodologia utilitzada en aquell moment.

Això darrer és possible perquè s'ha optat per treballar sempre sobre dades ja publicades, sigui recentment o fa alguns anys. El conjunt d'algorismes són d'ús habitual o introduïts en aquestes publicacions (vegi's [4], [22], [2] o [5]) i, com ja s'ha anat exposant, els problemes de prova pertanyen al repositori UCI ([3]) o bé a repositoris propis, i han estat presentats també en les corresponents publicacions (vegi's [49], [54], [55] o [56]).

El títol de cada apartat es refereix a la novetat o evolució proposada en els algorismes assajats i, per fer-ne més fàcil la seva identificació, s'hi introdueix també la cita de la publicació corresponent on aquesta novetat ha estat presentada.

9.1 Anàlisi de la fase de recuperació en un CBR amb memòria clusteritzada [2]

El raonament basat en casos (CBR, [80]) conté diverses etapes sobre les quals s'han proposat força modificacions o millores, en funció del problema al qual s'aplica. En d'altres publicacions del nostre Grup ([5] o [21]) s'ha treballat sobre la fase de recuperació, tot intentant fer-la més eficient a partir de la clusterització de la memòria de casos: habitualment, la recuperació es realitza tenint en compte tota aquesta memòria, i el procés d'anàlisi per clústers hauria de permetre uns resultats com a mínim equivalents, amb un cost computacional menor.

També en d'altres treballs ([22]) s'ha estudiat l'efecte que la complexitat inherent a cada problema de prova té sobre la bondat de l'algorisme que s'hi assaja. La conclusió, exposada àmpliament en el capítol 4, és que no és possible un anàlisi curós de la bondat dels algorismes sense tenir el compte la complexitat dels problemes de prova sobre els quals s'assaja. Més encara: convé fer el citat anàlisi en funció de la regió de complexitat (concepte també introduït a 4) a la qual pertany cada problema de prova, doncs els resultats poden ser considerablement diferents per a cada regió.

Donat aquest context, els resultats publicats a [2], que s'analitzaran a continuació, intenten determinar una metodologia general per permetre a un expert una anàlisi global de l'efecte de la clusterització de la memòria de casos, tenint en compte l'efecte de la complexitat dels problemes de prova i el fet de trobar-se davant un problema multivariant: la precisió en la classificació i el cost computacional del procés, lligat directament amb el número d'operacions necessàries per a la fase de recuperació en el CBR, són les magnituds que en determinaran la bondat.

Els problemes de prova analitzats són un total de 56, que es reparteixen per regions de complexitat segons es veu a la taula 5.1: 17 són de baixa complexitat, 30 de mitjana complexitat i 9 d'alta complexitat. Tots ells són problemes de dues classes, o bé han estat manipulats per ser-ho. Com ja s'ha dit en el capítol 4, això facilita l'anàlisi de la seva complexitat inherent.

Els algorismes assajats són un total de 13, d'entre els quals un d'ells coincideix amb la metodologia clàssica del CBR (l'anomenat *All_All*). La nomenclatura utilitzada, que es pot veure classificada a la figura 9.1 segons el número de clústers del SOM seleccionats i la quantitat d'elements de cada un d'ells que s'utilitzen per a la fase de recuperació, es defineix més àmpliament a [2].

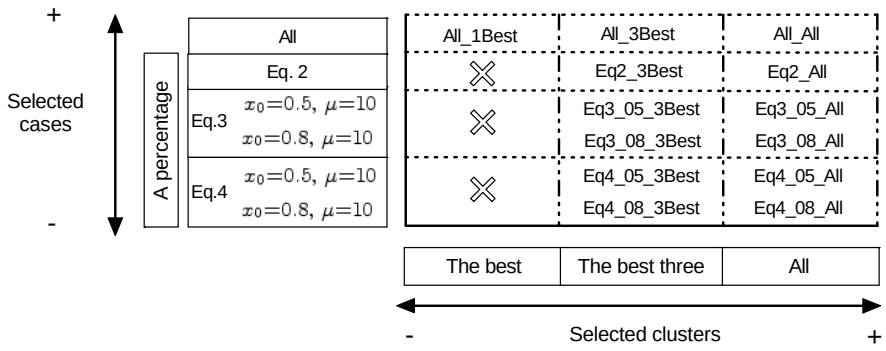

Figura 9.1: Esquema que defineix les configuracions dels 13 algorismes assajats. Les vconfiguracions marcades amb una creu no s'han utilitzat, perquè són equivalents a la *All_1Best*.

Durant el treball s'han utilitzat ja alguns dels resultats que aquí s'analitzaran: en concret, als apartats 5.2, 6.2.3, 6.3.2 i 7.4.1. Com ja s'ha exposat, aquests han estat resultats parcials, que a més només afecten a problemes de prova d'unes determinades regions de complexitat. El que es farà a continuació és l'anàlisi complert seguint el proposat a l'apartat **??**, separant l'anàlisi segons la complexitat dels problemes de prova. Degut a l'extensió que es requeriria, no s'inclou la taula amb els resultats per als 13 algorismes i els 56 problemes de prova: correspon a la taula 5.2 i la figura 5.4, utilitzades precisament com un exemple en què la representació gràfica proposada pot estalviar una immensa taula de resultats. Tots aquests resultats s'han obtingut a partir d'un *10-folds cross-validation* amb estratificació.

9.1.1 Problemes de prova de baixa complexitat

Per als 17 problemes de baixa complexitat, el càlcul de l'estadístic F de Fisher permet assegurar l'existència d'una diferència significativa entre alguns dels 13 algorismes assajats, sense necessitat de comprovar prèviament el compliment de les condicions d'aplicabilitat de l'anàlisi de variàncies, doncs el valor obtingut és un ordre de magnitud superior al valor crític per $\alpha = 0.05$: a la taula 9.1 es pot veure com $F >> F_{0.05}$.

Si es pensa en termes del valor de p, resulta bastant evident que un

SS_T	2238.2
SS_{BA}	89.0
SS_{BD}	2055.8
SS_{res}	93.4
df_{BA}	12
df_{BD}	16
df_{res}	192
MS_{BC}	7.42
MS_{BS}	128.48
MS_{res}	0.49
Estadístic F	15.25
$F_{crit}(\alpha = 0.05)$	1.80
$F_{crit}(\alpha = 0.01)$	2.28

Taula 9.1: Anàlisi de variàncies pels 13 algorismes assajats, sobre els problemes de prova de complexitat baixa. El valor de l'estadístic $F >> F_{0.05}$ permet assegurar l'existència d'una diferència significativa entre els algorismes.

hipotètic incompliment de les condicions d'aplicabilitat no és crític: el valor de F obtingut equival a un $p \simeq 10^{-22}$. D'acord amb això, podem assegurar l'existència d'alguna diferència significativa al nivell marcat per $\alpha = 0.05$, més enllà que alguna de les condicions d'aplicabilitat de l'anàlisi de variàncies no es compleixi.

A partir d'aquí, cal determinar quines seran les comparacions que es realitzaran. Tenint en compte que una de les estratègies és la clàssica del CBR (A_1) i la resta són modificacions introduïdes a partir de la clusterització de la memòria de casos ($A_2, ..., A_{13}$), l'objectiu serà estudiar com són les variacions observades entre aquests i el CBR. En primer lloc, plantegem la hipòtesi nul·la per la qual les modificacions introduïdes per la clusterització no aporten, en conjunt, una diferència de comportament significativa respecte A_1. És a dir, es realitza el càlcul del contrast de $X_2, ..., X_{13}$ respecte X_1, amb una hipòtesi nul·la que cerca la seva igualtat:

$$H_0 : X_1 = \frac{X_2 + ... + X_{13}}{12} \tag{9.1}$$

El valor de F obtingut en aquest càlcul és també clarament superior al crític per $\alpha = 0.05$, com es veu a la taula 9.2, que utilitza les definicions descrites a l'apartat 7.3.3.

A continuació, un cop demostrat que l'anterior hipòtesi nul·la és rebutjable, es realitza el càlcul dels diferents valors possibles de la distància crítica

SS_{comp}	11.89
MS_{comp}	11.89
MS_{res}	0.49
df_{comp}	1
df_{res}	192
F	24.45
$F_{.95}$	3.9
$F_{.99}$	6.77

Taula 9.2: Valors obtinguts per al càlcul del contrast de la hipòtesi nul·la entre l'estratègia clàssica del CBR i la resta d'algorismes assajats, per als problemes de prova de complexitat baixa.

CD, per concloure quins algorismes tenen, respecte el CBR, un comportament significativament diferent.

Els resultats es mostren a la taula 9.3, i el seu anàlisi segons l'exposat a l'apartat 7.3.3 es pot veure a la taula 9.4, on s'estudia en quins casos existeixen diferències significatives. En aquest problema, i tenint en compte que tots els algorismes mostren un valor inferior de precisió en la classificació respecte el CBR (o, el que és el mateix, un valor major de l'error comès), aquests resultats porten a concloure que 9 dels 12 algorisme tenen un comportament significativament pitjor que el CBR, mentre que els altres 3 (que correspon als resultats X_3, X_7 i X_9) tenen un comportament equivalent.

$F_{df_{comp},df_{res},.05}$	3.89
$\alpha_{PC} = 0.05/12$	0.0042
$t_{B\|D,.05}$	2.9
CD_{LSD}	0.35
$CD_{B\|D}$	0.51

Taula 9.3: Valors obtinguts per al càlcul de la distància crítica CD, seguint les metodologies LSD de Fisher i Bonferroni-Dunn, per als problemes de prova de complexitat baixa.

Una representació d'aquest resultat es veu a la figura 9.2, on es representen els rangs mitjans de cada algorisme, i els valors de CD a partir del rang del clàssic CBR. Aquells algorismes amb una diferència de rang respecte el CBR menor que CD_{LSD} tenen un comportament significativament equivalent, mentre que aquells amb diferència superior a $CD_{B|D}$ tenen un comportament significativament diferent. En aquest cas, no hi ha cap algorisme que tingui una diferència de rang entre CD_{LSD} i $CD_{B|D}$.

	Dif.	$Dif > CD_{B\|D}$	$Dif < CD_{LSD}$
$X_2 - X_1$	1.21	Si	
$X_3 - X_1$	0.33		Si
$X_4 - X_1$	1.30	Si	
$X_5 - X_1$	1.84	Si	
$X_6 - X_1$	0.61	Si	
$X_7 - X_1$	0.08		Si
$X_8 - X_1$	0.69	Si	
$X_9 - X_1$	0.12		Si
$X_10 - X_1$	0.97	Si	
$X_11 - X_1$	1.38	Si	
$X_12 - X_1$	1.60	Si	
$X_13 - X_1$	1.82	Si	

Taula 9.4: Valors de la diferència respecte el resultat obtingut pel CBR (X_1). En aquells casos en què aquesta és superior a $CD_{B|D}$ es pot afirmar que existeix una diferència significativa, mentre que quan la diferència és menor que CD_{LSD}, es pot afirmar just el contrari.

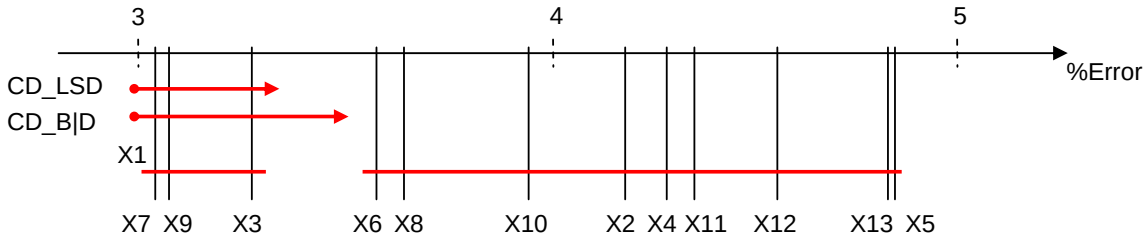

Figura 9.2: Representació gràfica dels valors de l'error mitjà per a cada algorisme, per als problemes de prova de complexitat baixa. Els valors de CD calculats estan representats a partir del valor mitjà de X_1, que correspon al clàssic CBR. Aquells algorismes amb una diferència de valors respecte el CBR menor que CD_{LSD} tenen un comportament significativament equivalent, mentre que aquells amb diferència superior a $CD_{B|D}$ tenen un comportament significativament diferent.

9.1.2 Problemes de prova de complexitat mitjana

D'igual manera que amb l'anterior grup, pels 30 problemes de complexitat mitjana es pot calcular el valor de l'estadístic de Fisher, obtenint un valor clarament superior al valors crític i, per tant, afirmant l'existència d'una diferència significativa sense comprovar les condicions d'aplicabilitat de l'anàlisi de variàncies. A la taula 9.5 es mostren els resultats obtinguts, on es veu de nou com $F >> F_{0.05}$.

SS_T	31201.5
SS_{BA}	650.3
SS_{BD}	28812.4
SS_{res}	1738.8
df_{BA}	12
df_{BD}	29
df_{res}	348
MS_{BC}	54.19
MS_{BS}	993.53
MS_{res}	5.00
Estadístic F	10.85
$F_{crit}(\alpha = 0.05)$	1.78
$F_{crit}(\alpha = 0.01)$	2.24

Taula 9.5: Anàlisi de variàncies pels 13 algorismes assajats, sobre els problemes de prova de complexitat mitjana. El valor de l'estadístic $F >> F_{0.05}$ permet assegurar l'existència d'una diferència significativa entre els algorismes.

A continuació, es determinen les comparacions a estudiar. D'igual manera que en el cas anterior, es comença amb una hipòtesi de contrast respecte el resultat obtingut pel CBR, que a nivell de precisió és el millor dels 13 algorismes, i és el control respecte el qual es vol comparar a la resta. De manera idèntica als problemes de baixa complexitat, el valor de l'estadístic F calculat indica l'existència d'una diferència significativa entre el CBR i la resta algorismes en conjunt (veure taula 9.6), resultat que ve confirmat en el càlcul de la distància crítica CD (taula 9.7).

Aquest darrer resultat es representa gràficament a la figura 9.3, on es pot veure com els algorismes A_3, A_7 i A_9 mostren un comportament significativament equivalent al del A_1. Els resultats són, malgrat la diferència en la complexitat dels problemes de prova, similars als tinguts per als problemes

SS_{comp}	55.83
MS_{comp}	55.83
MS_{res}	5.00
df_{comp}	1
df_{res}	348
F	11.18
$F_{.95}$	3.87
$F_{.99}$	6.71

Taula 9.6: Valors obtinguts per al càlcul del contrast de la hipòtesi nul·la entre l'estratègia clàssica del CBR i la resta d'algorismes assajats, per als problemes de prova de complexitat mitjana.

$F_{df_{comp},df_{res},.05}$	3.87
$\alpha_{PC} = 0.05/12$	0.0042
$t_{B\|D,.05}$	2.88
CD_{LSD}	0.84
$CD_{B\|D}$	1.23

Taula 9.7: Valors obtinguts per al càlcul de la distància crítica CD, seguint les metodologies LSD de Fisher i Bonferroni-Dunn, per als problemes de prova de complexitat mitjana.

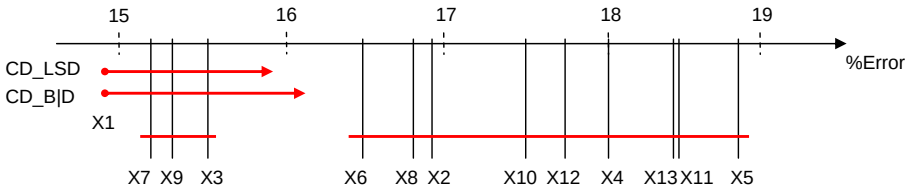

Figura 9.3: Representació gràfica dels valors de l'error mitjà per a cada algorisme, per als problemes de prova de complexitat mitjana. Els valors de CD calculats estan representats a partir del valor mitjà de X_1, que correspon al CBR clàssic. Aquells algorismes amb una diferència de valor respecte el CBR menor que CD_{LSD} tenen un comportament significativament equivalent, mentre que aquells amb diferència superior a $CD_{B|D}$ tenen un comportament significativament diferent.

de complexitat baixa. Caldrà veure, el següent apartat, si això es complexi també per als problemes d'alta complexitat o si, com s'havia previst, l'anàlisi porta a resultats diferents en funció de la complexitat inherent als problemes de prova.

9.1.3 Problemes de prova d'alta complexitat

Seguint l'esquema referit anteriorment, es desenvolupen en primer lloc els càlculs necessaris per al valor de F sobre els 9 algorismes de complexitat elevada, obtenint aquesta vegada un valor que no permet rebutjar H_0 sense preocupar-se de les condicions d'aplicabilitat de l'anàlisi de variàncies, ans al contrari: el valor obtingut és menor al crític i, per tant, cal estudiar el compliment de les condicions per estar segurs d'assumir la conclusió obtinguda, que seria la no existència de diferència significativa entre els 13 algorismes assajats. Els resultats es mostren a la taula 9.8, amb $F < F_{0.05}$.

SS_T	15040.2
SS_{BA}	16.8
SS_{BD}	14573.6
SS_{res}	449.7
df_{BA}	12
df_{BD}	8
df_{res}	96
MS_{BC}	1.40
MS_{BS}	1821.70
MS_{res}	4.68
Estadístic F	0.30
$F_{crit}(\alpha = 0.05)$	1.85
$F_{crit}(\alpha = 0.01)$	2.38

Taula 9.8: Anàlisi de variàncies pels 13 algorismes assajats, sobre els problemes de prova d'alta complexitat. El valor de l'estadístic $F < F_{0.05}$ no permet assegurar l'existència d'una diferència significativa entre els algorismes.

La primera condició a comprovar, seguint el protocol definit a l'apartat 7.5, és la possible violació de la normalitat de les dades. Es tractaria d'assegurar que els resultats obtinguts per qualsevol de les diferències $X_i - X_j$ en cap cas s'ajusten a un comportament bimodal i, per tant, que no es pot

rebutjar d'entrada la hipòtesi de normalitat. S'ha de tenir en compte que, amb només 9 problemes de prova, les dades no són suficients com per aplicar els habituals algorismes de comprovació de normalitat. En aquest sentit, i d'acord amb la prudència que s'ha recomanat al llarg del capítol 5, una sola distribució $X_i - X_j$ propera a la bi-modalitat ja serà motiu suficient com per posar en dubte la viabilitat d'un anàlisi de variàncies, i es proposarà un test no paramètric.

El número de comparacions possibles per parelles de M algorismes és igual a $M(M-1)/2$, i per tant convé determinar quines de les diferències possibles entre dos algorismes són més susceptibles de violar la gaussianitat de manera ràpida. Una possibilitat és calcular la matriu de covariàncies dels resultats dels M algorismes, i escollir aquelles parelles que corresponen a valors de la covariància més elevats. Si això es du a terme pel problema que ens ocupa, es troba que les comparacions amb covariància més elevada són entre els algorismes A_2 i A_{11}, entre l'A_2 i l'A_5, i entre l'A_1 i l'A_{11}. Només amb el primer d'aquests casos, l'histograma de les diferències ja mostra una distribució de baixa normalitat (veure figura 9.4), fet que aconsella no considerar el resultat de l'anàlisi de variàncies i continuar amb un anàlisi no paramètric.

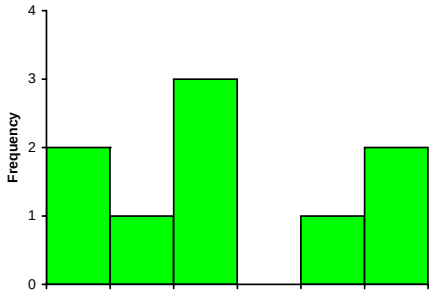

Figura 9.4: Histograma de la diferència dels resultats pels algorismes A_2 i A_{11}. Les poques dades disponibles i la forma obtinguda no permeten afirmar amb rotunditat que no hi ha una distribució amb comportament bi-modal.

Seguint el protocol presentat a l'apartat 7.5, l'alternativa és utilitzar una anàlisi no paramètrica, que s'inicia amb un test de Friedman sobre el conjunt de les dades, amb el qual es posa a prova la hipòtesi nul·la segons la qual no existiria cap diferència significativa entre el comportament dels M algorismes. Les suposicions per aplicar el test de Friedman són menors que les necessàries per fer el mateix amb l'anàlisi de variàncies: tan sols és necessari que la mostra de N problemes de prova hagi estat seleccionada a l'atzar

d'entre la població de problemes de prova existents, i que de les mesures disponibles es pugui treure informació ordinal, essent originalment una variable contínua ([109]). És a dir, que permeti ordenar-les de major a menor per cada problema de prova. Ambdues es compleixen clarament en aquest cas.

D'acord amb la metodologia exposada a l'apartat 7.4.1, s'obtenen els resultats de la taula 9.9, que permeten rebutjar la hipòtesi nul·la (doncs $\chi^2_{F,cor} > \chi^2_{F,.05}$ i $F_I > F_{I,.05}$) i, per tant, afirmar que entre els 13 algorismes hi ha alguna diferència significativa.[1]

	$M = 13,\ N = 9$
χ^2_F	39.18
C	1.00
$\chi^2_{F,cor}$	39.18
df	12
$\chi^2_{F,.05}$	21.03
Rebuig H_0	Si
F_I	4.55
df	12 x 96
$F_{I,.05}$	1.85
Rebuig H_0	Si

Taula 9.9: Resultats de l'aplicació del test de Friedman sobre els 13 algorismes per als 9 problemes de prova de complexitat alta. Els valors de l'estadístic obtinguts permeten rebutjar l'opció nul·la, per qualsevol de les dues metodologies possibles d'aplicació del test.

El següent pas és escollir la manera de realitzar les comparacions a posteriori, tenint en compte que ara el millor algorisme no és el CBR clàssic, sinó el corresponent a S_OAN_05, A_7 segons la nomenclatura utilitzada amb aquestes dades. Aquets fet s'observa directament a la figura 5.4, ja comentada anteriorment.

Amb el resultat d'aquest algorisme com a control respecte el qual es compara el comportament de la resta, s'aplica un test de Bonferroni-Dunn per al càlcul de $CD_{B|D}$, seguint l'exposat a 7.4.2. El resultat que s'obté és

$$CD_{B|D} = 5.23 \tag{9.2}$$

[1] Cal destacar que, en tant que es tracta d'una anàlisi no paramètrica, a partir d'aquest moment el rang substitueix al percentatge d'error de classificació com a mesura de la bondat de l'algorisme.

i, d'acord amb la representació gràfica de la figura 9.5, porta a concloure que existeix un nombrós grup d'algorismes amb comportament no significativament diferent que A_7, mentre que n'hi ha tres (A_{12}, A_5 i A_{13}) que es pot assegurar tenen un comportament significativament pitjor. Aquest resultat ve confirmat per un test de Holm sobre totes aquestes dades respecte A_7, com es pot veure a la taula 9.10.

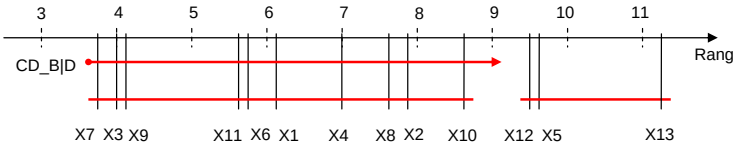

Figura 9.5: Representació gràfica dels rangs mitjans de cada algorisme, per als problemes de prova de complexitat alta. El valor de $CD_{B|D}$ calculat està representat a partir del rang mitjà de A_7, que correspon a l'algorisme amb un millor comportament per aquests problemes de prova. Aquells algorismes amb una diferència de rang respecte el mínim menor que $CD_{B|D}$ tenen un comportament significativament equivalent, mentre que aquells amb diferència superior a $CD_{B|D}$ tenen un comportament significativament diferent.

9.1.4 Resum i conclusions

El resum de la metodologia emprada per a l'anàlisi dels resultats publicats a [2] es mostra a l'esquema de la taula 9.21, on també s'hi exposen els resultats obtinguts. En primer lloc, cal destacar el fet que les conclusions varien depenent de la complexitat dels problemes de prova sobre els quals s'assajen els diferents algorismes analitzats. Aquest fet posa de manifest la necessitat, ja exposada al capítol 4, d'incloure l'estudi de la complexitat del problema de prova en l'anàlisi de la bondat d'un algorisme.

En segon lloc, es veu també clarament com els mateixos test estadístics no són vàlids per qualsevol dels anàlisis realitzats: depenent dels problemes de prova sobre els quals s'assajen els algorismes, només uns determinats test són possibles d'aplicar, doncs no es compleixen les condicions necessàries per a l'aplicabilitat d'uns altres. Sempre que sigui possible, com ja s'ha esmentat repetidament, s'aplicaran test paramètrics, com es fa pels problemes de complexitat baixa i mitjana. Quan no sigui així, es continua pel camí no paramètric marcat a l'esquema del protocol a seguir (veure figura 7.5).

Finalment, l'anàlisi permet afirmar que existeixen un conjunt d'estratègies de clusterització de la memòria de casos que aporten resultats equiva-

i	**Alg. comp.**	$R_i - R_0$	z_i	p_i	$\frac{\alpha}{(M-i)}$
12	$S\_EBN\_M\_3$	0,17	0,09	0,397	0.0500
11	S_OAN_08	0,22	0,12	0,396	0.0250
10	$S\_OAN\_05\_N$	1,89	1,03	0,235	0.0167
9	$S\_OAN\_05\_M\_3$	2	1,09	0,22	0.0125
8	CBR	2,22	1,21	0,192	0.0100
7	$S\_PEBN\_M\_3$	3,17	1,72	0,09	0.0083
6	$S\_OAN\_08\_M\_3$	3,89	2,12	0,042	0.0071
5	S_OBM	4,11	2,24	0,033	0.0063
4	$S\_OAN\_05\_M\_3\_N$	4,72	2,57	0,015	0.0056
3	$S\_OAN\_08\_M\_3\_N$	5,67	3,09	0,003	0.0050
2	S_PEBN	5,72	3,12	0,003	0.0045
1	$S\_OAN\_08\_N$	7,39	4,02	< 0.001	0.0042

Taula 9.10: Aplicació del mètode de Holm sobre els resultats dels problemes de prova de complexitat alta. Els valors obtinguts, a partir de la comparació amb l'A_7, permeten rebutjar la hipòtesi d'igualtat de comportament pels 3 algorismes amb valor de p menor. Com de costum, el valor de confiança utilitzat és $\alpha = 0.05$.

lents als del CBR i independentment de la complexitat dels problemes de prova: és el cas dels algorismes A_3, A_7 i A_9, que corresponen a les estratègies $S\_EBN\_M\_3$, S_OAN_05 i S_OAN_08. Aquest conjunt d'algorismes s'amplia amb els A_{11}, A_8, A_1, A_4, A_2, A_{10} i A_6 quan la complexitat dels problemes és elevada. En aquest darrer cas, hi ha fins a 5 estratègies amb un resultat de rang mitjà que millora el del CBR.

A banda, hi ha un grup format pels algorismes A_5, A_{12} i A_{13} (que equivalent a les estratègies S_PEBN, $S\_OAN\_08\_M\_3\_N$ i $S\_OAN\_08\_N$) que en qualsevol cas mostren uns resultats significativament pitjors que la millor estratègia per cada cas.

Aquests darrers resultats responen a la hipòtesi plantejada a l'inici de l'estudi que es va publicar a [2]: existeix un grup de tres algorismes, fruit de determinades estratègies de clusterització de la memòria de casos, pels quals no es perd precisió respecte el CBR i, en canvi, tenen necessitat d'un número menor d'operacions en la fase de recuperació de la memòria de casos, fruit de la pròpia clusterització. Per tant, es conclou que és possible reduir el cost computacional del CBR, mitjançant la clusterització de la memòria de casos, sense una pèrdua de precisió significativa.

	Classe A	Classe B	Classe C
$F >> F_{0.05}$	Si	Si	No
Bi-modalitat	-	-	No
Esfericitat	-	-	-
scd	-	-	-
H_0 ANOVA	Rebuig	Rebuig	-
Contrast	Rebuig	Rebuig	-
Anàlisi CD	A_3, A_7, A_9 equivalents al CBR	A_3, A_7, A_9 equivalents al CBR	-
H_0 Friedman	-	-	Rebuig
Holm (control)	-	-	A_5, A_{12} i A_{13} sign. pitjors que CBR

Taula 9.11: Esquema dels resultats de la comparació dels 13 algorismes sobre els 56 problemes de prova, separats per regions A, B i C de complexitat (baixa, mitjana i alta, respectivament).

9.2 Millores de la representació ADI en un LCS amb Algorismes Genètics sota enfocament de Pittsburgh ([4])

Els algorismes genètics (GA, [137]) s'han aplicat sovint per a generar classificadors que resolguin problemes com els plantejats fins ara en aquest treball ([138], [139]). Una de les tipologies més utilitzades són els coneguts com a LCS (*learning classifier systems*), amb dos principals enfocaments: Pittsburgh ([140]) i Michigan ([138]).

En aquests enfocaments, la representació del coneixement acostuma a ser a través de dades nominals, per la qual cosa es necessita un algorisme de discretització per transformar els valors reals dels atributs, tractant els intervals generats com a valors nominals. Un bon algorisme de discretització és aquell que manté un compromís entre la pèrdua d'informació deguda a què s'agrupen valors diferents en un mateix interval, i l'augment desmesurat d'aquests intervals. Per a fer-ho, una bona opció és la coneguda com les regles de representació ADI (*adaptive discretization intervals*, [51]), utilitzada en l'enfocament de Pittsburgh.

L'article de Bacardit i Garrell analitzat ([4]) treballa diverses variacions d'aquest ADI, a partir de combinar les millores proposades en el propi article, relatives a la proporció de cada discretitzador en la població i el número

d'instàncies per atribut. L'objectiu de l'article és demostrar com alguna d'aquestes variacions proposades millora el comportament del classificador que en resulta, a partir de les dades obtingudes de l'assaig sobre una col·lecció de problemes de prova. Aquests resultats són comparats amb l'ADI original, i també amb el resultat d'altres classificadors de tipologia molt diferent, com el C4.5 ([52]) i el IB1 ([53]), amb la qual cosa es comparen un total de 8 algorismes, a partir dels resultats obtinguts de l'assaig sobre una col·lecció de 15 problemes de prova.

Les dades que s'utilitzen per intentar demostrar aquesta millora s'han presentat ja a la taula 7.9, i s'han obtingut a partir d'un *10-folds cross-validation* amb estratificació, amb 15 iteracions. Tant en el capítol 6 (relacionant-ho amb les matrius de guanys allí definides) com en el capítol 7 (exemplificant l'ús del test de Friedman per a discutir una hipòtesi nul·la), s'han utilitzat ja aquests resultats. El que es farà en aquest apartat és desenvolupar l'anàlisi sencer de les dades, i veure fins a quin punt les conclusions obtingudes pels autors, i publicades en l'article original, són certes.

9.2.1 Anàlisi paramètric

D'acord amb les metodologies exposades en els capítols 4 i 5, el primer que caldria fer, per proposar una hipòtesi nul·la, seria estudiar les propietats inherents dels problemes de prova sobre els quals s'assagen els algorismes classificadors, amb l'objectiu de descartar el que passava en l'exemple mostrat en l'apartat anterior: que els problemes siguin de tipologies molt diferents, i els classificadors mostrin un comportament diferent depenent de la tipologia dels problemes de prova.

En aquest problema, però, apareix un primer fet que ho dificulta: els algorismes han estat assajats sobre una col·lecció de problemes en què n'hi ha que tenen més de dues classes, i és precisament sobre aquests que es vol discutir la seva bondat. Ja es va comentar al capítol 4 que les mètriques de complexitat presentades es podien calcular per problemes de dues classes, i per tant no es podria repetir el mateix anàlisi en aquest cas.

No obstant això, no sembla que els problemes sobre els quals s'assagen els algorismes tinguin en aquest cas diferències inherents gaire importants, com a mínim que afectin a la bondat d'aquests algorismes significativament. Si fos el cas, hi hauria un grup de problemes pels quals uns algorismes determinats obtindrien molt millors resultats que uns altres, i aquest mateix comportament no es trobaria en la resta dels problemes. Això hauria de produir un efecte sobre la dispersió del resultat que s'obté per cada problema de prova.

Aquesta opció es pot descartar amb un simple anàlisi de la relació de la mitjana dels resultats obtinguts per cada problemes de prova, i la dispersió d'aquests resultats: com es pot observar a la figura 9.6, la relació és altament lineal i, per tant, no apareixen els fenòmens comentats anteriorment. Podem afirmar que la comparació entre els algorismes es pot realitzar sense problemes sobre tota la col·lecció de problemes de prova que es mostra a la citada taula 7.9. És a dir, que considerar tota la col·lecció de problemes no amagarà conclusions, pel fet que hi hagi comportaments que només es mostrin en un tipus de problema de prova.

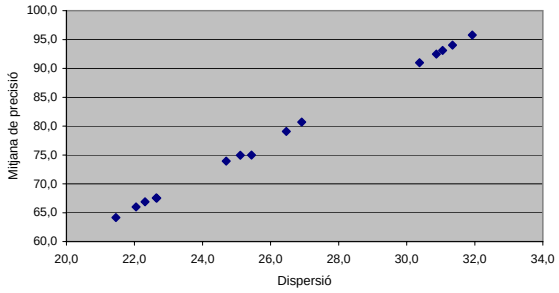

Figura 9.6: Relació entre el valor mig dels vuit algorismes proposats sobre cada un dels problemes de prova i la dispersió d'aquesta mateixa mesura. Cada punt de la gràfica representa el resultat d'un problema de prova. Les dades originals es mostren a la taula 7.9, i en la figura s'aprecia clarament la relació lineal entre ambdues magnituds.

Un cop resolta aquesta qüestió, convé ara realitzar l'estudi paramètric en si mateix, començant per la discussió de si es pot aplicar o no l'anàlisi de variàncies (ANOVA), d'acord amb el que descriu l'esquema que es pot veure a la figura 7.1. En primer lloc, es fa el càlcul de l'estadístic F, segons el definit a l'apartat 7.3.1, obtenint-se els valors de la taula 9.12, amb $F < F_{crit}$.

El valor de F obtingut no permet afirmar que es pugui aplicar l'anàlisi de variàncies, doncs no compleix que $F >> F_{crit}$, ans al contrari. D'acord amb l'esquema abans referit, el següent pas és comprovar que no es viola la normalitat de les dades, les diferències obtingudes entre totes les parelles possibles de resultats. Com es comparen 8 algorismes, es podrien arribar a construir 28 diferències: per estudiar un dels casos que pot ser "pitjor", es calcula la matriu de covariàncies i s'estudien aquelles diferències que estan més correlades.

En el cas que ens ocupa, tots els elements de la matriu de covariàncies són molt propers a 1, la qual cosa indica una elevada correlació entre les mesures obtingudes, cosa que ja es pot observar directament sobre la taula de resul-

Variable	Valor
SS_T	15620.60
SS_{BA}	4.83
SS_{BD}	15120.57
SS_{res}	495.34
df_{BA}	7
df_{BD}	14
df_{res}	98
MS_{BC}	0.69
MS_{BS}	1080.03
MS_{res}	5.05
Estadístic F	0.14
$F_{crit}(\alpha = .05)$	2.10
$F_{crit}(\alpha = .01)$	2.83

Taula 9.12: Anàlisi de variàncies per als resultats obtinguts de l'assaig dels 8 algorismes proposats sobre els 15 problemes de prova. S'observa com el valor de F trobat no permetria rebutjar la hipòtesi H_0 (doncs $F < F_{crit}$), en cas que es complissin les condicions per poder aplicar l'anàlisi de variàncies.

tats: a nivell de la precisió de classificació, és bastant més important sobre quin problema de prova es discuteix, que quin algorisme hi ha estat aplicat. Això porta a què, independentment de quina parella de resultats s'estudia, la normalitat de les dades queda suficientment garantida: un exemple és el que es mostra a la figura 9.7, on es mostra clarament que no es podrà afirmar que les dades segueixen una distribució de tipus bi-modal i, a nivell del que requereix el protocol resumit a la figura 7.1, és suficient com per no descartar encara l'anàlisi de variàncies.

A continuació, cal mirar si es compleixen les condicions d'esfericitat, tot aplicant un test com el de Bartlett ([7], introduït a l'apartat 7.3.2). Els resultats obtinguts, que es mostren a la taula 9.13, no permeten assegurar el compliment d'aquestes condicions (al seu moment ja es va destacar la dificultat que això arribés a passar en un problema real), i per tant convé estudiar la simetria composta dèbil. Aquest darrer concepte s'avalua a partir del model de Myers i Well ([8]), obtenint-se les dades que apareixen a la taula 9.14. Els resultats d'aquest darrer test tampoc permeten assegurar que es compleixi la suposició de simetria composta dèbil i, per tant, cal optar per utilitzar un estudi no-paramètric.

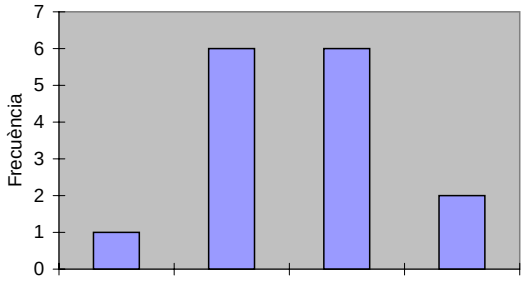

Figura 9.7: Histograma de la diferència de resultats entre els algorismes ADI3 i ADI4. Malgrat les poques dades disponibles, sembla evident que no es pot rebutjar l'aplicació de l'anàlisi de variàncies basant-se en un suposat comportament bi-modal de les dades. La figura obtinguda és similar a les que es podrien trobar per qualsevol altre parell d'algorismes, en aquest problema.

Variable	**Valor**
χ^2	171.6
df	8
p	<0.001
Esfricitat	No

Taula 9.13: Anàlisi de la suposició d'esfericitat per les dades obtingudes per cada un dels 8 algorismes. Com mostra el valor obtingut per p, es pot rebutjar la hipòtesi nul·la sobre el compliment de l'esfericitat. El test que s'aplica per obtenir el resultat és el de Bartlett ([7]).

Variable	**Valor**
Var_{MAX}	180.11
Var_{min}	126.22
df	13
t	4.47
$t_{.05}$	2.16
scd	No

Taula 9.14: Anàlisi del compliment de la suposició simetria composta dèbil (scd) per les dades obtingudes pels 8 algorismes comparats. Seguint les indicacions de Myers i Well ([8]), no es pot assegurar el compliment d'aquesta suposició.

9.2.2 Anàlisi no-paramètric

L'esquema de la figura 7.1 indica que els resultats obtinguts ens porten a una anàlisi no-paramètrica dels resultats, d'acord amb l'esquema de la figura 7.5, que incorpora també la primera. Seguint el que indica, el primer que cal fer és calcular l'estadístic F i comparar-ho amb el valor crític per l'α corresponent, i discutir la hipòtesi nul·la del problema, que indicaria que no existeix diferència de comportament entre els 8 algorismes comparats, i que per tant les diferències obtingudes serien fruit estrictament de variacions aleatòries.

Aquest càlcul s'ha realitzat ja a l'apartat 7.4, com el primer exemple de l'aplicació del test de Friedman sobre un conjunt de dades, amb els diferents estadístics que es poden calcular ($\chi^2_{F,cor}$ i F_I). Els resultats obtinguts, que es reprodueixen de nou a la taula 9.15, indiquen que no és possible rebutjar aquesta hipòtesi nul·la global plantejada i que, per tant, amb les dades obtingudes no es pot afirmar l'existència de diferències significatives entre els 8 algorismes del problema. Seguint l'esquema de la figura 7.5, amb aquesta afirmació finalitza l'anàlisi d'aquestes dades.

Variable	**Valor**
χ^2_F	12.42
C	0.96
$\chi^2_{F,cor}$	12.89
df	7
$\chi^2_{F,.05}$	14.07
Rebuig H_0	No
F_I	1.88
df	7 x 98
$F_{I,.05}$	2.10
Rebuig H_0	No

Taula 9.15: Resultats de l'aplicació del test de Friedman sobre els resultats de la taula 7.9. Els valors de l'estadístic obtinguts no permeten rebutjar l'opció nul·la.

Veient aquest resultat, i sabent que ja era conegut d'abans (càlculs de l'apartat 7.4), es podria argumentar que tots els càlculs realitzats en l'apartat anterior sobre el compliment de les condicions per a l'aplicació de l'anàlisi de variàncies (normalitat, esfericitat, simetria composta) no eren necessaris. Això és cert, doncs un cop s'obté un valor de $F < F_{crit}$ i un test de Friedman en el mateix sentit ($\chi^2_{F,cor} < \chi^2_{F,.05}$), es pot afirmar que no es podrà rebutjar la hipòtesi nul·la: tant si es pot aplicar l'anàlisi de variàncies com si no, la

conclusió és la mateixa. No obstant això, i d'acord amb l'objectiu del capítol, s'han volgut desenvolupar tots els apartats del càlcul previst al protocol proposat al capítol 7.

9.2.3 Resum i conclusions

L'anàlisi estricte de les dades obtingudes ha portat, doncs, a concloure que aquestes no permeten afirmar l'existència de diferències significatives entre els 8 algorismes comparats al problema publicat a l'article de Bacardit i Garrell ([4]). Per tant, no es podria afirmar allò que els autors volien: que les millores introduïdes en els algorismes ADI portaven a una millora real en la capacitat de classificació, mesurada per l'encert en la fase de test.

Aquest resultat obtingut té algunes implicacions molt importants, a nivell de la comparació de la bondat d'un conjunt d'algorismes tal i com s'estudia en aquest treball. En primer lloc, i com ja s'ha exposat a l'apartat 7.4, això contradiu els resultats obtinguts en la comparació simple per parelles que s'ha fet a l'apartat 6.3.4, quan s'ha introduït la matriu de guanys, tot intentant extrapolar a la comparació múltiple tècniques que només són vàlides per a la comparació d'un algorisme respecte un altre.

En la discussió de la matriu de guanys, s'ha acabat afirmant que l'algorisme ADI4 és "el millor de tots, amb 72 guanys per només 27 casos en què es comporta pitjor", d'igual manera que s'ha dit que, en la comparació per parelles, "l'algorismes $ADI4$ és significativament millor" que els algorismes ADI, $ADI1$, $ADI3$ i $ADI5$. Cap d'aquestes conclusions és certa: serien vàlides si només es coneguessin les dades necessàries per a la comparació simple de dos d'aquests algorismes, però en el moment en què es coneixen els resultats de més de dos, les metodologies per parelles deixen de tenir validesa.

En el cas que ens ocupa, per exemple, la informació aportada per tots els resultats impedeix rebutjar la hipòtesi nul·la global, segons la qual no hi ha diferències en el comportament dels 8 algorismes comparats: i, si no es pot rebutjar la hipòtesi nul·la global, no té cap sentit discutir sobre comparacions particulars (veure el propi esquema de la figura 7.5). És aquest un exemple més del mal ús de les metodologies de comparació simple en un problema de comparació múltiple.

Finalment, cal destacar també que les conclusions obtingudes no tenen res a veure amb les que els autors varen publicar: l'article afirma que la configuració ADI4 "*is a clear winner*" i que "*is better (in average) than C4.5 and IB1*". També diu que la configuració ADI "*has a robust and reliable*

behaviour". A la vista dels càlculs fets en el present apartat, s'observa com les conclusions publicades no coincideixen amb les obtingudes aquí, a banda del fet que la manera d'expressar-les no és correcte: no té cap significat, pel cas que ens ocupa, dir que un algorisme és millor "*in average*" que un altre.

De ben segur, la llargada i complexitat dels càlculs que cal realitzar, segons les metodologies proposades en aquest treball, no faciliten la discussió estricte dels resultats obtinguts, perquè és molt més simple l'elaboració d'una matriu de guanys, per exemple. Tot i així, en un exemple com el present ha quedat demostrat el risc d'extreure conclusions sense una anàlisi ben acurada dels resultats obtinguts. A la taula 9.16 es mostra el resum dels càlculs que han permès arribar a les conclusions exposades.

$F >> F_{0.05}$	No
Bi-modalitat	No
Esfericitat	No
scd	No
H_0 ANOVA	-
Contrast	-
Anàlisi CD	-
$\chi^2_{F,cor} > \chi^2_{F,crit}$	No
H_0 Friedman	Acceptació

Taula 9.16: Esquema dels resultats de la comparació dels 8 algorismes sobre els 15 problemes de prova. El no compliment de les condicions d'esfericitat i de simetria composta dèbil (scd) impedeixen l'ús d'un test paramètric (ANOVA), i el valor de $\chi^2_{F,cor}$, obtingut pel test de Friedman, porta a acceptar la hipòtesi nul·la.

9.3 Introducció d'una funció de pertinença al SOMCBR per a reduir-ne l'error de classificació ([5])

La variació del SOMCBR que serà analitzada en aquest apartat és especialment útil com a eina de suport per al diagnosi mèdic. Els casos estudiats estan relacionats amb el càncer de mama, considerada la principal de les dolències de càncer entre les dones de 15 a 54 anys. L'estudi és possible a partir de l'anàlisi d'imatges mamogràfiques, que permeten l'estudi del teixit mamari i la detecció primerenca de les formacions i estructures fora del normal ([141]).

Les imatges que en resulten no són fàcilment analitzables, i encara menys fàcil és extreure'n una conclusió clara sobre possibles alteracions observades. Com a eina de suport, s'han generat en els darrers anys el que es coneixen com a *Computer Aided Systems*, amb l'objectiu de facilitar l'ús de grans bancs d'informació incerta com la que resulta de les mamografies. Aquests sistemes, com el DESMAI (*Decision Support System Based on Mammographic Images*, [50], [142]), aporten informació a l'expert i l'ajuden en el procés de la presa de decisió.

En el cas concret que s'analitza, el sistema ha vingut desenvolupat a partir d'un esquema de raonament basat en casos (CBR), sobre el qual s'han implantat tècniques de clusterització de la memòria de casos ([21]), per tal de reduir l'espai de cerca en l'etapa de recuperació (*retrieval*). Sobre aquest esquema, s'han introduit diverses variants, en funció de:

- Quin és el sistema pel qual es decideix la classe de cada nou element: per votació entre els veïns més propers sense clusteritzar la memòria de casos (CBR), per votació entre els veïns més propers clusteritzant la memòria de casos ($SOMCBR-vot$), o bé introduint una funció de pertinença i un nivell llindar en la discussió dels veïns més propers amb la memòria clusteritzada ($SOMCBR-per$).

- Quants veïns més propers (k-NN) són considerats per prendre la decisió: 1 (1-NN), 3 (3-NN) o 5 (5-NN).

Aquestes opcions, que estan extensament detallades a l'article citat al mateix títol de l'apartat ([5]), duen a poder considerar tres estratègies diferents (CBR, $SOMCBR-vot$ o $SOMCBR-per$), per les quals cal escollir entre tres valors de k-NN: 1-NN, 3-NN o 5-NN. Hi ha, per tant, 9 algorismes diferents per analitzar. El resultat de la seva aplicació porta a un problema multivariant, en què no tan sols cal considerar la precisió en la classificació com a mesura de bondat, sinó que també cal tenir en compte el percentatge d'elements no classificats: especialment en el cas del $SOMCBR-per$, el fet d'establir la pertinença a una classe o altra per una funció de pertinença i un nivell llindar, provoca que en alguns casos no sigui possible la classificació en una de les classes del problema (en aquest cas, benigne / maligne).

Els problemes de prova utilitzats són CA i BI ([54], [49]), DD, M3 i MB ([55], [56]). Els tres darrers casos s'han separat en problemes de dues classes (amb el procediment que s'ha mostrat a l'apartat 4.4) per tenir un conjunt de problemes de prova majors, i poder analitzar els valors de les mètriques de complexitat. A la taula 9.17 es mostren les seves principals

característiques, i a la taula 9.18 es reprodueixen els resultats obtinguts, pel que fa al percentatge d'error de classificació comès i al percentatge d'elements no classificats[2]. També s'hi inclou el percentatge de reducció en el número d'operacions que cal realitzar a la fase de recuperació respecte el necessari pel CBR ($\%R$): un dels efectes de la clusterització és la reducció de l'espai de cerca i, per tant, la reducció amb el número d'operacions necessàries per a classificar el nou cas.

	Prob. prova	**Atr.**	**Ins.**	**Clas.**	**Distribució per classes**
BI	Biopsia	25	1027	2	0 (530), 1 (497)
CA	Mamografia	22	216	2	benign (121), malign (95)
DD	DDSM	143	501	4	b1(61), b2(185), b3(157), b4(98)
M3	MIAS-3C	153	322	3	fatty(106), dense(112), glandular(104)
MB	MIAS-Birads	153	320	4	b1(128), b2(78), b3(70), b4(44)

Taula 9.17: Descripció dels problemes de prova utilitzats en el treball publicat a [5]. De cada un d'ells s'indica l'habitual abreviació, el nom complert, el número d'atributs de cada instància (contant-hi la classe a la qual pertanyen), el número d'instàncies total, el número de classes, i la distribució de les instàncies per cada classe.

En els capítols anteriors, s'han utilitzat aquests problemes de prova i aquests resultats diverses vegades. Per exemple, en l'apartat 3.3, on s'ha discutit sobre l'error de classificació, s'ha fet servir aquest cas per mostrar una manera de treballar que pot tenir sentit en aquells problemes on el cost d'un error sigui molt elevat: variar la manera d'efectuar la pròpia classificació, tot renunciant a la classificació en aquells casos més "dubtosos", pot provocar un descens de l'error, tot i que a canvi augmentin els elements no classificats. És precisament el que mostra la taula 9.18.

En el capítol 5 ha aparegut dues vegades aquest problema: primer, en l'apartat de l'anàlisi de variàncies (7.3.1), s'ha estudiat l'efecte sobre la precisió de l'augment del valor de k, comprovant el compliment de les condicions que determinen el domini d'ús d'aquest test; després, en l'apartat 7.4.1, s'hi ha aplicat el test de Friedman en aquella configuració que no complia les condicions per a l'aplicació de l'anàlisi de variàncies. En els següents apartats es completaran tots aquests càlculs i es realitzarà l'anàlisi sencer del problema.

[2]Com en la majoria d'exemples utilitzats en aquest treball, els resultats s'han obtingut a partir d'un *10-folds cross-validation* amb estratificació

	1-NN		3-NN		5-NN		%R
	%Err.	%No Class.	%Err.	%No Class.	%Err.	%No Class.	
CBR-vot							
CA	37.5	0.0	35.2	0.0	33.3	0.0	-
BI	16.8	0.0	17.3	0.0	16.0	0.0	-
DD c1	17.0	0.0	15.2	0.0	15.6	0.0	-
DD c2	33.7	0.0	34.1	0.0	30.7	0.0	-
DD c3	32.1	0.0	31.0	0.0	32.7	0.0	-
DD c4	22.2	0.0	20.0	0.0	19.4	0.0	-
MB c1	12.2	0.0	10.3	0.0	12.2	0.0	-
MB c2	24.1	0.0	22.5	0.0	20.9	0.0	-
MB c3	21.2	0.0	16.6	0.0	16.9	0.0	-
MB c4	9.7	0.0	9.7	0.0	10.6	0.0	-
M3 c1	19.9	0.0	22.1	0.0	22.4	0.0	-
M3 c2	12.1	0.0	9.6	0.0	10.2	0.0	-
M3 c3	25.8	0.0	27.9	0.0	28.3	0.0	-
SOMCBR-vot							
CA	37.0	0.6	35.6	0.6	35.3	3.3	53.8
BI	23.8	0.0	21.6	0.0	21.5	0.0	86.8
DD c1	17.7	0.0	15.7	0.0	14.4	0.0	87.4
DD c2	38.1	0.1	37.4	0.2	37.1	0.1	86.9
DD c3	36.8	0.0	36.6	0.0	35.7	1.6	86.2
DD c4	25.7	0.0	22.3	0.0	21.5	0.1	83.9
MB c1	20.2	0.3	20.4	0.3	21.3	0.5	86.3
MB c2	30.2	0.1	26.1	0.1	26.1	0.1	84.3
MB c3	26.6	0.2	24.4	0.2	26.1	0.1	84.3
MB c4	13.9	0.1	12.9	0.1	12.9	1.0	80.8
M3 c1	27.3	0.3	26.5	0.3	28.1	1.0	87.2
M3 c2	15.7	0.1	17.1	0.1	17.3	0.9	83.5
M3 c3	32.8	0.1	33.0	0.1	31.8	1.5	85.1
SOMCBR-per γ=0.6							
CA	38.6	3.3	23.8	60.1	18.5	78.1	53.8
BI	24.0	0.0	10.2	43.8	7.1	60.6	86.8
DD c1	18.4	0.0	9.7	24.6	7.4	35.3	87.4
DD c2	37.7	0.1	25.9	61.6	22.6	79.5	86.9
DD c3	36.7	1.6	26.3	52.8	17.3	71.6	86.2
DD c4	25.6	0.1	14.5	37.8	17.3	71.6	83.9
MB c1	20.1	0.5	9.6	90.9	8.0	61.9	86.3
MB c2	31.3	0.1	15.6	87.1	10.9	69.2	84.3
MB c3	26.7	0.2	14.1	88.5	8.0	58.3	84.8
MB c4	14.4	1.0	7.6	92.8	6.8	31.4	80.8
M3 c1	27.3	1.3	16.9	68.8	10.0	68.7	87.2
M3 c2	16.5	1.0	9.0	88.9	7.1	50.9	83.5
M3 c3	32.7	1.5	24.1	42.6	19.4	72.9	85.1

Taula 9.18: Resultats del percentatge d'error de classificació (%Err.), i del percentatge de casos no classificats (%No Class.), per les tres estratègies assajades (on γ indica el valor llindar per la funció de pertinença del $SOMCBR-per$), i pels diferents valors de $K-NN$: 1-NN, 3-NN i 5-NN. La taula també inclou el percentatge de reducció en el nombre d'operacions necessàries a l'etapa de recuperació, en comparació amb les necessàries pel CBR sense clusterització.

9.3.1 Estudi dels problemes de prova

En primer lloc, d'acord amb el comentat al capítol 4, es procedeix a l'estudi de les propietats inherents als problemes de prova. En aquest cas, en tractar-se de l'aplicació de diverses variants del CBR, es pot utilitzar el mapa de complexitat ja definit a l'apartat 4.7, i publicat prèviament a l'article [22]: un esquema bi-dimensional amb les mètriques $F3$ i N_{12}. En aquest cas, l'estudi és possible perquè els algorismes han estat assajats sobre els problemes de la taula 9.17 convertits en problemes de dues classes, com s'observa a la taula de resultats 9.18.

L'estudi dels problemes des d'aquest punt de vista duu a un resultat com el que es mostra a la figura 9.8. En aquesta figura, cada punt correspon al valor de $(F3, N_{12})$ per un problema de prova, i hi estan representats els d'aquest problema i la resta dels 56 utilitzats a [2] (i que estan descrits a la taula 5.1). Es pot observar com, excepte el punt corresponent al problema *miasbi*2*c*4, la resta es troben en una mateixa zona del mapa de complexitat.

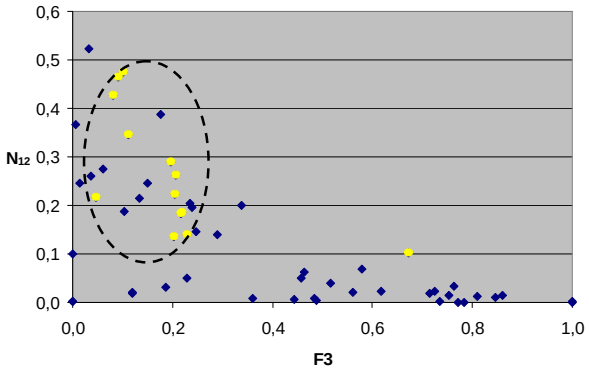

Figura 9.8: Mapa de complexitat $(F3, N_{12})$, on s'hi representen tots els 56 problemes utilitzats a [2]. Els símbols grocs representen els 13 problemes on s'han assajat els algorismes estudiats, amb els resultats mostrats a la taula 9.18. S'observa com, excepte el problema *miasbi*2*c*4, al qual correspon el punt de color groc amb $F3 \in (0.6, 0.7)$, la resta es troben en una regió de complexitat similar.

En el sentit de la complexitat, es podrà considerar al problema *miasbi*2*c*4 com un *outlier*. Per tant, els resultats de la discusió de les hipòtesis es farà amb i sense aquest problema, per tal de valorar-ne l'efecte. Excepte aquest cas particular, però, els valors de les mètriques per als problemes que es tracten ens indiquen que no caldrà fer una anàlisi d'hipòtesis diferenciada per grups de problemes de prova.

9.3.2 Relació entre k i l'error de classificació

Un cop determinat que la tipologia dels problemes de prova no requereix un tractament diferenciat d'alguns d'ells, es procedeix a analitzar si els canvis proposats en els algorismes afecten positivament a la seva bondat, entesa en aquest cas com una reducció en l'error de classificació.

Tal i com s'ha dit abans, en aquest cas també caldrà tenir en compte altres factors (com el percentatge d'elements no classificats, o la reducció en el número d'operacions a realitzar), però això serà analitzat en el següent apartat, en tant que porta a un problema multivariant. Aquestes seran les variables que diferenciaran el comportament de les tres estratègies proposades (CBR, $SOMCBR-vot$, $SOMCBR-per$), i per cada una d'elles el determinant és veure com afecta al resultat la variació en k, que determina el número de veïns propers que s'utilitza a la fase de recuperació.

Per cada una d'aquestes tres estratègies, hi ha tres valors possibles de k, i per tant es tracta d'un problema de comparació múltiple. Seguint l'esquema proposat a la figura 7.5, el primer que cal fer és calcular l'estadístic F de l'anàlisi de variances, cosa que porta a uns resultats com els que es mostren a la taula 9.19. En un dels casos (CBR), el valor de F i F_{crit} no permet rebutjar la hipòtesi nul·la a un nivell de singificança d'$\alpha = 0.05$. En el cas de l'estratègia $SOMCBR-vot$ sí que seria possible, però no es compleix el fet que $F >> F_{crit}$.

Per tant, aquests dos casos demanarien seguir amb l'anàlisi segons el protocol proposat, que fa necessari l'estudi de les condicions del domini d'ús de l'anàlisi de variàncies. En canvi, el valor de F trobar per $SOMCBR-per$ permet assegurar que el rebuig de la hipòtesi nul·la és ben fiable, doncs $F >> F_{crit}$ (com, d'altra banda, no és difícil de suposar veient els resultats obtinguts, taula 9.18).

La citada taula 9.19 també mostra com el problema de prova que té consideració d'*outlier*, des del punt de vista de les mètriques de complexitat, no afecta qualitativament les conclusions que s'obtenen, sobre el rebuig o acceptació de la hipòtesi nul·la. Per tant, es seguirà l'anàlisi sense tenir-lo en compte d'una manera especial. Tot i això, és important tenir present que la presència d'alguns problemes de prova més en la regió de complexitat del *miasbi*2*c*4 podria afectar les conclusions extretes, doncs els valors de F serien considerablement diferents per un i altre grup de problemes: la sola presència d'aquest problema afecta al valor de F en un percentatge superior al 10% del seu valor, en els tres casos estudiats.

En l'apartat 7.3.1 s'ha discutit prèviament el compliment de les condi-

	CBR		**SOMCBR-vot**		**SOMCBR-per**	
	(outl.)	**(no outl.)**	**(outl.)**	**(no outl.)**	**(outl.)**	**(no outl.)**
SS_T	2688.76	3634.21	2447.05	3937.82	3252.97	3984.7
SS_{BA}	10.25	11.12	11.99	10.54	1493.94	1371.8
SS_{BD}	2632.25	3578.29	2407.59	3898.96	1608.82	2375.13
SS_{res}	46.26	44.82	27.48	28.32	150.21	237.77
df_{BA}	2	2	2	2	2	2
df_{BD}	12	12	12	12	12	12
df_{res}	24	24	24	24	24	24
MS_{BC}	5.12	5.56	5.99	5.27	746.97	685.9
MS_{BS}	219.35	298.19	200.63	324.91	134.07	197.93
MS_{res}	1.93	1.87	1.15	1.18	6.26	9.91
Estadístic F	2.66	2.98	5.23	4.46	119.35	69.23
$F_{crit}(\alpha = 0.05)$	3.40	3.40	3.40	3.40	3.40	3.40
$F_{crit}(\alpha = 0.01)$	5.61	5.61	5.61	5.61	5.61	5.61

Taula 9.19: Anàlisi de variàncies per cada una de les tres estratègies assajades a [5]. Com mostren els valors obtinguts per l'estadístic F, en dos dels tres casos es poden rebutjar les hipòtesi nul·les, H_0. També s'observa com les conclusions no varien si no es té en compte el problema *miasbi2c4* (casos indicats com "not outl.").

cions que determinen si els resultats obtinguts permeten aplicar l'anàlisi de variàncies o no. A les taules 7.3 i 7.4 s'han mostrat els resultats per a les condicions d'esfericitat i de simetria composta dèbil, que porten a les conclusions esquematitzades a la figura 7.2: mentre que pel cas del CBR es pot aplicar l'anàlisi de variàncies i, per tant, es pot afirmar directament que l'augment de k no provoca diferències significatives en l'error de classificació per aquesta estratègia, pel cas del $SOMCBR - vot$ no es compleixen les condicions d'aplicació.

En aquest darrer cas, cal optar per una alternativa no paramètrica: a la taula 9.20 s'exposen els resultats obtinguts d'aplicar el test de Friedman sobre els tres valors possibles de k per a l'estratègia $SOMCBR - vot$, i el resultat no ens permet afirmar l'existència de diferències significatives. Tal i com s'ha exposat a l'apartat 7.4.1, aquest resultat és una prova més de la poca capacitat que el test de Friedman té per posar de manifest diferències significatives, més a prop del que seria el test binomial que no pas el de Wilcoxon, si es fa una comparativa amb les comparacions simples ([19]).

Aquest exemple ens situa en un cas extrem, fruit del to conservador del protocol proposat a l'apartat 7.3.2, en què es discuteix el domini d'aplicabilitat de l'anàlisi de variàncies: les condicions no són suficients per aplicar-lo

	(outl.)	**(no outl.)**
χ^2_F	5.69	4,86
C	1	1
$\chi^2_{F,cor}$	5.69	4,86
df	2	2
$\chi^2_{F,.05}$	5.99	5,99
Rebuig H_0	No	No
F_I	3.36	2,77
df	3 x 13	3 x 12
$F_{I,.05}$	3.40	3,44
Rebuig H_0	No	No

Taula 9.20: Resultats de l'aplicació del test de Friedman sobre els resultats de l'estratègia $SOMCBR - vot$ introduida a [5], els resultats de la qual es mostren a la taula 7.1. Els valors de l'estadístic obtinguts no permeten rebutjar l'opció nul·la, i l'abscència de l'outlier no faria sinó reforçar aquesta conclusió.

(quan la seva major potència, en el sentit definit a l'apartat 8.1, permetria rebutjar H_0), i el test no-paramètric no té prou capacitat per fer-ho, malgrat situar-se en valors de F_I propers al crític $F_{I,.05}$. En el seu moment ja s'ha fet evident que alguns autors, en aquest cas, proposarien directament el rebuig d'H_0 i l'aplicació del test a posteriori, però la tesi d'aquest treball és més conservadora, i mantindria la reticència al rebuig d'H_0. Si fos el cas que *realment* existís una diferència significativa com per rebutjar H_0, de ben segur apareixeria ràpidament en un test no-paramètric tot afegint algun problemes de prova a la col·lecció utilitzada.

El to conservador d'aquesta proposta també permet obtenir unes conclusions més robustes a la presència d'*outliers*, com és aquest cas: els resultats de la taula 9.20 mostren fins a quin punt la presència d'un sol problema en una regió de complexitat diferent pot arribar a modificar el resultat, fins a portar-lo al límit de canviar la conclusió que s'obtindria. És un exemple més de com de fonamental resulta fer una anàlisi prèvia de les propietats inherents dels problemes de prova.

9.3.3 Estudi del problema multivariant

De l'apartat anterior es conclou que, a un nivell de significança d'$\alpha = 0.05$, només en el cas de l'estratègia $SOMCBR - per$ l'augment del valor de k

aporta un guany significatiu en la reducció de l'error. Aquesta conclusió és correcte si es planteja el problema com, estrictament, una qüestió d'estudi de l'error de classificació. Si aquest és el cas, una simple observació de la taula dels resultats (9.18) permet veure com l'estratègia $SOMCBR - per$ és clarament millor que les altres, especialment per $k = 5$. Qualsevol anàlisi d'inferència aportaria aquest resultat amb un nivell de confiança molt elevat.

Ara bé, ja s'ha comentat que hi ha altres factors que també influeixen sobre el que s'entén per la bondat o comportament dels algorismes discutits: el número d'operacions a realitzar a la fase de recuperació i el percentatge de casos no classificats són variables a ser tingudes en compte, amb major o menor pes en funció de quin sigui l'objectiu principal de les modificacions introduïdes als algorismes.

En general, caldria definir una funció de cost que depengués de les tres variables del problema (per això es parla d'un problema multivariant), i que amb l'ajust d'uns coeficients pogués donar una idea de quina és la bondat *real* de l'algorisme en qüestió. Si s'anomena x a l'error de classificació, y al percentatge de casos no classificats i z a la reducció del número d'operacions que cal realitzar, una possibilitat seria l'ús d'una funció de cost f definida com

$$f(x, y, z) = \alpha x + \beta y + \gamma(1 - z) \tag{9.3}$$

on els coeficients α, β i γ ajusten la importància de cada variable en el valor final d'aquesta funció de cost, per la qual valors propers a zero indicarien un comportament òptim: error baix de classificació, pocs elements no classificats i elevada reducció del temps de càlcul. El cas particular tractat a l'apartat anterior correspon al cas límit en què $\alpha = 1$ i $\beta = \gamma = 0$.

Les característiques de cada problema determinaran la forma de la funció f i els valors dels corresponents coeficients. El cas que s'estudia a continuació considera les variables error de classificació (x) i percentatge de casos no classificats (y), per poder comparar el comportament en funció de k: les magnituds relatives al temps de càlcul depenen només de l'aplicació o no d'un mapa auto-organitzat (SOM) sobre la memòria de casos i, per tant, el resultat és com a totes les variants SOM estudiades, independent del valor de k.

A la figura 9.9 es mostren els resultats, per aquestes dues variables, en el cas de l'estratègia $SOMCBR - per$. Cada punt de la figura representa els resultats d'un problema de prova, i es pot observar com els valors obtinguts permeten una certa agrupació en funció del valor de k (de fet, l'escala logarítmica en el percentatge de l'error de classificació pretén mostrar més clarament aquest efecte). Es veu com l'augment del valor de k implica un augment dels

no classificats (per $k = 1$ aquest és pràcticament nul), però també una reducció de l'error de classificació. Són, per tant, tendències contradictòries en el sentit de la bondat de l'algorisme. La definició de la funció $f(x, y)$ determinaria quin pes es dóna a cada una d'aquestes tendències i, finalment, quina configuració serà preferible.

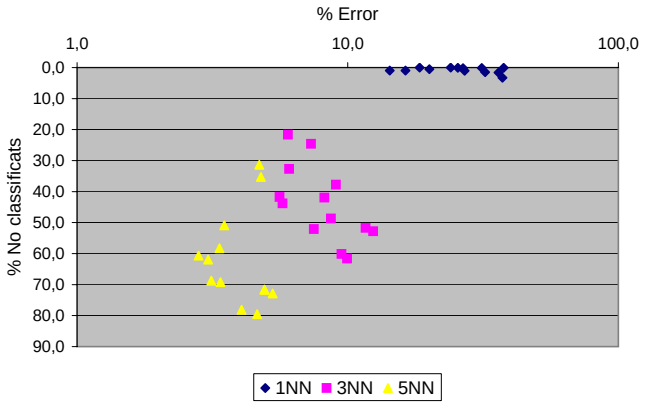

Figura 9.9: Representació dels resultats obtinguts per als 13 problemes de prova, en un esquema representat per les variables error de classificació i percentatge d'elements no classificats. Cada punt és el resultat d'un problema de prova, i l'eix logarítmic s'utilitza per mostrar millor les diferents agrupacions de les dades.

Els resultats obtinguts també permeten un darrer anàlisi, d'acord amb els valors de les mètriques de complexitat representats a la figura 9.8. Al capítol 4 s'han exposat els resultats publicats a l'article [22], que han permès determinar un mapa de complexitat especialment adient en l'estudi d'un CBR amb la memòria clustritzada per un procés SOM. D'aquells raonaments semblava poder concloure's que la complexitat del problema de prova, expressada segons $F3$ i N_{12}, tenia una certa influència sobre la bondat resultant de l'algorisme assajat, per les diferents variants tipus SOMCBR. Aquesta relació es pot observar també en els resultats obtinguts en aquest problema de la següent manera.

D'entrada, suposem que la bondat de l'algorisme és màxima per un resultat que se situï al punt $(1, 0)$ de la gràfica 9.9 (és a dir, nul error de classificació, i tots els casos classificats). D'acord amb això, es podria definir una funció de cost que vingués donada per la distància euclidiana a aquest punt ($dist_{XY}$, sumant-hi els valors dels tres possibles valors de k), de tal manera que la bondat de l'algorisme és major per valor menors d'aquesta funció (que, a més, es pot normalitzar entre els valors de 0 i 1 per a major

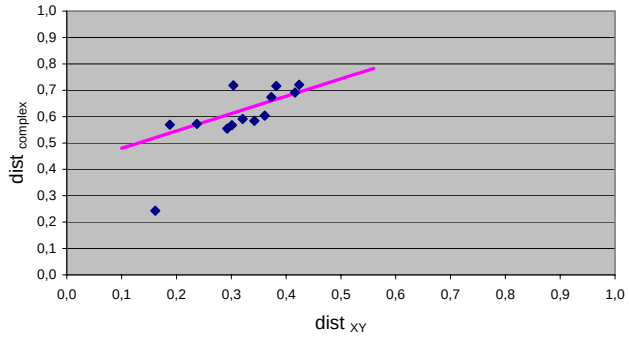

Figura 9.10: Representació desl valors que, per cada problema de prova, prenen les magnituds definides com $dist_{XY}$ i $dist_{complex}$. Els valors propers al punt $(0,0)$ indicarien mínima complexitat i màxima bondat de l'algorisme assajat. S'hi observa una relació de linealitat, excepte pel cas del problema de prova *miasbi*2*c*4, que ja s'ha comentat que es pot considerar un *outlier* degut als valors de les seves propietats inherents, mesurades a través de les mètriques de complexitat.

comoditat).

A banda, sobre el mapa de complexitat representat a la figura 9.8 es pot definir també una distància euclidiana respecte el punt $(1,0)$, que en aquest cas representa la mínima complexitat possible. Aquesta nova distància, que anomenem $dist_{complex}$, també es calcula normalitzada entre 0 i 1. Doncs bé, a la figura 9.10 es mostra la relació entre ambdues distàncies, i s'hi observa una relació lineal per al conjunt de problemes de prova, excepte pel cas del *miasbi*2*c*4, que un cop més mostra el seu caràcter d'*outlier* des del punt de vista de les mètriques de complexitat.

Aquesta relació indica com, efectivament, existeix una certa proporcionalitat entre la complexitat del problema de prova i la bondat de l'algorisme, per a l'estratègia $SOMCBR-per$: com menor és la complexitat del problema ($dist_{complex} \rightarrow 0$), major és la bondat de l'algorisme ($dist_{XY} \rightarrow 0$). Un cop més, aquest exemple referma les conclusions del capítol 4, i els resultats publicats a [22].

9.3.4 Resum i conclusions

L'anàlisi dels resultats obtinguts ha portat a un problema de tipus multivariant, la qual cosa permet una discussió dels resultats des de diversos punts de vista. Els algorismes comparats han estat un total de 9 (tres possibilitats de clusterització-decisió, i tres possibilitats en quant als veïns propers a consid-

erar), i els resultats s'han obtingut sobre una col·lecció de 13 problemes de prova.

Un primer estudi dels problemes ha determinat que, excepte el *miasbi*2*c*4, la resta ocupen tots una regió de complexitat similar, amb la qual cosa no és necessari estudiar les hipòtesis per diferents grups de problemes. Com a molt, caldrà tenir en compte el caràcter d'*outlier* del *miasbi*2*c*4, i mirar l'efecte de no considerar-lo sobre les conclusions.

Posteriorment, l'estudi de l'efecte del número de veïns propers a tenir en compte en la fase de recuperació (k), ha dut a la conclusió que només pel cas del $SOMCBR - per$ l'evolució d'aquest paràmetre implica diferències significatives en els resultats obtinguts, si es pren l'error de classificació com a única mesura de la bondat dels algorismes. A la taula 9.21 s'esquematitzen els càlculs realitzats per arribar a aquesta conclusió.

	CBR	**SOMCBR-vot**	**SOMCBR-per**
$F >> F_{0.05}$	No	No	Si
Bi-modalitat	No	No	-
Esfericitat	No	No	-
scd	Si	No	-
H_0 ANOVA	Acceptació	-	Rebuig
$\chi^2_{F,cor} > \chi^2_{F,crit}$	-	No	-
H_0 Friedman	-	Acceptació	-

Taula 9.21: Esquema dels resultats de la comparació dels 3 valors de k possibles ($k = 1$, $k = 3$ i $k = 5$) per cada una de les tres estratègies assajades (CBR, $SOMCBR - vot$ i $SOMCBR - per$), a partir de les dades obtingudes sobre els 13 problemes de prova. Les dades sobre la possible bi-modalitat, esfericitat i scd ja es tenien del capítol 7, i les seves implicacions han donat lloc a la figura 7.2.

Finalment, s'ha procedit a estudiar breument el problema des d'un punt de vista multivariant, doncs són fins a 3 les variables implicades: l'error de classificació, el percentatge d'elements no classificats i la reducció del número d'operacions de la fase de recuperació. L'anàlisi de les dues primeres, fet de manera combinada amb les mesures de complexitat calculades al capítol 4, han aportat una mostra més de l'efecte de les propietats dels problemes de prova sobre la bondat dels algorismes, en aquest cas per les variants del SOMCBR analitzades.

Capítol 10

Conclusions i línies de futur

10.1 Conclusions

La tesi que s'ha presentat s'emmarca dins l'àmbit del *machine learning* (ML): en concret, en aquells problemes en què una nova proposta d'algorisme d'aprenentatge és comparada amb aquells algorismes ja existents i, per fer-ho correctament, es requereix d'una metodologia experimental que permeti realitzar aquesta comparació, tenint en compte tots aquells elements necessaris pel fet de ser un problema d'inferència estadística.

La proposta realitzada cerca respondre als interrogants bàsics amb què una persona de l'àmbit es pot trobar en avaluar un nou algorisme d'aprenentatge, destinat a desenvolupar una certa tasca (classificació, predicció, etc.): de quina manera s'avalua la bondat d'aquest algorisme? Es pot avaluar o només fer-ne una estimació? Com es pot comparar aquesta bondat respecte la dels altres algorismes que desenvolupen la mateixa tasca? Com assegurar que el test d'inferència escollit per fer la comparació és l'adequat, i quina magnitud ho mesura? Quins efectes té sobre tot això la col·lecció de problemes de prova utilitzada per a estimar la bondat de l'algorisme?

La voluntat inicial d'aquest treball era respondre a aquestes preguntes, des d'un punt de vista teòric, però també tenint present l'aplicació als problemes habituals amb què es troba la comunitat del ML, i proposar un conjunt de metodologies que permetessin una anàlisi acurada del comportament d'un nou algorisme, passant per totes les etapes que han quedat reflectides a la figura 1.1.

Els treballs publicats fins al moment tenien, al nostre entendre, diverses mancances per a respondre a totes les qüestions plantejades: o bé es centraven

únicament en alguns aspectes molt concrets, oblidant-se d'altres igualment essencials per a garantir la validesa de la conclusió ([12], [13]); o bé eren estudis estrictament teòrics de difícil aplicació als casos pràctics ([38], [40]); o bé en l'intent d'adaptar-ho en l'àmbit del ML eren poc considerades qüestions teòriques importants, que portaven a simplificar en excés les conclusions ([9]).

Totes aquestes mancances han ajudat, possiblement, a que l'anàlisi del resultat en una publicació en aquest àmbit sigui habitualment parcial, i posaven de relleu la necessitat d'un treball que aportés solucions globals al problema. D'aquest plantejament inicial n'ha sorgit aquesta tesi.

La tesi s'ha iniciat amb el capítol 2, on s'han presentat els antecedents, terminologia i exemples que s'utilitzen al llarg del treball. Ja en el capítol 3 s'han respost les primeres preguntes, en aquest cas sobre la bondat d'un algorisme [1]. Aquesta bondat ha de ser avaluada a partir d'un estimador que es desitjarà amb mínim biaix i variància. La conclusió exposada és que, malgrat cap de les opcions possibles és òptima en tots els casos, els treballs permeten concloure que la iteració d'un *cross-validation* estratificat amb 10 *folds* garanteix sempre un biaix extremadament petit, mentre que la pròpia iteració redueix la variància a valors també suficientment assumibles.

Queda molt clar també el que en cap cas es pot realitzar: comparar mesures de diferents algorismes que, tot i que puguin significar el mateix[2], hagin estat obtingudes per estimadors diferents. En aquest cas el biaix provocat per aquests diferents estimadors sobre les mesures de bondat podria ser d'ordre de magnitud de les pròpies diferències existents i, per tant, provocar que s'arribés a una conclusió errònia.

També en aquest capítol 3 s'ha discutit la possibilitat d'augmentar les restriccions sobre el procés que realitza l'algorisme (per exemple, la classificació) amb l'objectiu de reduir considerablement l'error, en aquells casos en què quan es produeixi impliqui un cost molt elevat o inassumible (com un error de diagnosi en un problema mèdic). La conclusió, a partir de l'exemple estudiat tot restringuint el procés de recuperació en una variant del CBR, és que aquesta manipulació pot tenir un efecte molt beneficiós en el resultat, tot i implicar una reducció important en la capacitat operativa del propi algorisme, doncs hi ha molts casos sobre els quals no es pot concloure. La funció de cost, que atorgui a cada variable un pes en la determinació de la bondat de l'algorisme, permetria en cada cas l'anàlisi comparatiu.

[1] Bondat que s'estima en realitzar la tasca que té encomanada.

[2] Aquest raonament és vàlid independentment de la mesura de bondat utilitzada: l'error en la classificació, la precisió en el mateix procés, l'àrea sota la corba ROC, etc.

Un cop tractat com avaluar un algorisme, en el capítol 4 s'han estudiat els efectes que les propietats inherents dels problemes de prova sobre els quals s'estima la seva bondat tenen en la comparació entre algorismes. La primera conclusió és que no estudiar aquestes propietats pot amagar diferències significatives entre algorismes, si aquestes es presenten només per alguns tipus de problemes. La segona és que tot un conjunt de mètriques de complexitat, definides en aquest capítol, són una bona eina per construir les regions de complexitat introduïdes, que permeten separar els problemes en funció del comportament que sobre ells tindran els algorismes estudiats.

A més, l'estudi d'aquestes propietats també ha portat a la introducció d'uns esquemes gràfics que simplifiquen considerablement les interminables taules de resultat que apareixen habitualment als treballs. Aquests esquemes, definits al capítol 5, faciliten també l'anàlisi sobre quines comparacions estudiar, a partir del correcte plantejament de les hipòtesis del problema. Precisament en aquest capítol 5 s'han introduït els conceptes adients per plantejar aquestes hipòtesis.

En els capítols 6 i 7 s'ha estudiat el que és pròpiament la comparació entre els algorismes, separant els casos de comparació simple (un algorisme respecte un altre, capítol 6) dels casos de comparacions múltiples (un conjunt de més de dos algorismes, capítol 7). Per ambdós casos s'han estudiat amb profunditat les alternatives en quant als test d'inferència estadística per a realitzar la comparació, i s'ha proposat un protocol d'actuació.

Pel cas de comparacions simples, s'ha definit un protocol que permet estudiar el domini d'ús del t-test, incloent-hi la possibilitat de relaxar algunes de les restriccions que determinen la seva validesa, en funció de les dades del problema i les hipòtesis estudiades. Això ha portat a proposar un protocol d'aplicació dels test per comparacions simples (incloent-hi l'estudi dels test no paramètrics) i a definir un conjunt de condicions per a l'aplicació del que es coneix com a matrius de guanys per a les comparacions múltiples. La conclusió sobre aquesta darrera qüestió és que cal descartar, sempre que sigui possible, l'extrapolació de les tècniques de comparació simple a un problema de comparació múltiple a través d'aquestes matrius, perquè les conclusions que se n'extreuen sovint son errònies.

En el capítol 7 s'ha fet un plantejament similar aplicat als casos de comparacions múltiples. Tot l'estudi ha portat a la definició d'un protocol d'aplicació dels tests per a fer aquestes comparacions, incloent-hi una proposta d'ús del que es coneix com a distància crítica, per la qual existeixen diverses definicions, no sempre coincidents. La conclusió contradiu aquelles opinions segons les quals és preferible descartar d'entrada l'anàlisi de variàncies per

utilitzar directament test no paramètrics: el protocol proposat descriu exactament sota quines condicions són utilitzables els mètodes paramètrics[3], i en quins casos obligatòriament cal optar pels no paramètrics.

L'estudi de les condicions d'ús de cada un d'aquests test permet definir els protocols, que indiquen l'opció òptima en cada cas. De quina manera es pot analitzar, a posteriori de l'aplicació del test, si el test utilitzat seguint el protocol és realment l'òptim? Per a respondre a aquesta qüestió s'han plantejat algunes magnituds, descrites en el capítol 8, que aporten informació sobre la potència (definida com la capacitat d'un test per determinar l'existència d'una diferència significativa, en cas que aquesta existeixi) i la replicabilitat (definida com la probabilitat que un test repeteixi una conclusió si s'aplica repetidament).

A partir del càlcul en un problema de comparacions simples s'ha mostrat com les conclusions que s'obtenen (tot parametritzant i forçant les diferències entre els resultats obtinguts pels algorismes estudiats) són del tot coherents amb el protocol proposat: en aquells casos en que permet aplicar un test paramètric, els càlculs de la potència i la replicabilitat el consideren també com la millor opció. En canvi, quan cal optar per un test no paramètric perquè les condicions per l'ús d'un paramètric no es compleixen, aquest darrer apareix com el que pitjor resultat té segons les diferents variables relacionades amb aquests conceptes de potència i replicabilitat.

Per tant, es pot concloure que aquestes magnituds no cal que siguin estudiades si s'apliquen correctament els protocols proposats en aquest treball o, dit d'una altra manera, que el valor que prenguin són conseqüència directe de les condicions explicitades en aquests protocols.

Finalment, per facilitar la visió de conjunt de totes les propostes realitzades, i per veure quina és la seva aplicació concreta, es resolen diversos problemes plantejats a partir d'algorismes que el nostre Grup de Recerca ha presentat en publicacions fetes aquests darrers anys. La voluntat d'aquest darrer capítol també és convertir aquest treball en una eina més útil per a l'"usuari".

El conjunt de les propostes realitzades en aquesta tesi responen, doncs, les preguntes plantejades a l'inici del treball, quan l'objectiu era proposar una metodologia experimental per a l'anàlisi del comportament d'un nou algorisme d'aprenentatge, respecte d'un conjunt d'algorismes prèviament coneguts,

[3]Cal recordar un cop més que els mètodes paramètrics són el s que utilitzen major informació i per tant, si és possible la seva aplicació, sempre tindran un millor comportament, també en el sentit definit al capítol 8

sense entrar a fons en l'anàlisi del propi disseny d'aquest algorisme.

Aquesta metodologia experimental permet procedir amb un esquema d'actuació com el de la figura 1.1 per a l'anàlisi de resultats de sistemes d'aprenentatge artificial, amb la seguretat

1. que les mesures de bondat utilitzades seran òptimes (en el sentit d'un estimador de biaix i variància mínims),

2. que les propietats inherents dels problemes de prova no amagaran conclusions sobre la bondat dels nous algorismes sobre certs grups de problemes,

3. que les hipòtesis del problema estaran correctament plantejades (amb el corresponent control sobre la confiança del resultat),

4. i que la metodologia utilitzada per a la discussió de les hipòtesis estarà dins el seu domini d'aplicació i, per tant, que no només els resultats que se'n deriven seran fiables (al nivell de confiança determinat abans), sinó que de totes les opcions s'utilitzarà aquella amb major potència i replicabilitat.

10.2 Línies de futur

No obstant això, aquest no és un treball totalment tancat, sinó que permet obrir diverses línies de recerca futures, en bona part com a conclusió dels propis resultats obtinguts en aquest treball. Aquestes noves qüestions es podrien agrupar en les següents línies:

1. Justificacions teòriques d'algunes de les conclusions obtingudes.

   La primera línia d'actuació agruparia tots aquells casos en què l'experimentació en un o més problemes ens ha permès demostrat algunes qüestions, però sense aportar-hi una justificació teòrica concloent. Per exemple, un camí clar seria intentar trobar una relació genèrica entre la potència i la replicabilitat, i les restriccions per a l'aplicació dels test paramètrics, que donen forma als protocols proposats. Si els resultats trobats mostren coherència entre ambdues qüestions (resultats que també es podrien generalitzar per problemes de comparació múltiple), una línia de treball futura és establir el marc teòric que permeti justificar aquestes relacions.

2. Casos multivariants i funcions objectiu per determinar la bondat.

   Els casos multivariants són aquells en què més d'una magnitud determina la bondat de l'algorisme proposat; per exemple la precisió en la classificació i el seu cost computacional. En aquests casos fa falta encara un desenvolupament teòric que acabi amb protocols d'actuació de test d'inferència similar als exposats en aquesta tesi, basant-se també en la funció objectiu que determini el pes de cada variable a l'hora de fer la comparació. La determinació d'aquesta funció seria també un camp d'estudi interessant, en el cas concret del domini mèdic analitzat a l'apartat 9.3, per exemple. Aquesta seria una possible segona línia de futur, tenint en compte també la possibilitat de test com el MANOVA, esmentat a l'apartat 7.3.

3. Corbes operatives, corbes d'aprenentatge i aprenentatge incremental.

   Un tercer àmbit seria treballar l'aplicació d'aquestes metodologies quan la mesura de bondat és l'àrea sota la corba (AUC) de les corbes operacionals (ROC). En aquells casos en que la classificació ve determinada per una funció de pertinència i un llindar, podria aportar conclusions novedoses, com ja s'ha indicat al capítol 3. D'igual manera es podria treballar en les corbes d'aprenentatge discutides al mateix capítol, i la seva relació amb els casos en que l'algorisme té un procés d'aprenentatge incremental.

   Finalment, també caldria avaluar la potència i la replicabilitat dels test habituals si la mesura de bondat és l'àrea sota la corba ROC: d'una banda, no s'haurien de trobar diferències substancials amb el conclòs al capítol 8, i d'altra banda seria bo analitzar si les diferències existents es tornen més extremes o si, pel contrari, es redueixen.

4. Complexitat dels problemes de prova i bondat de l'algorisme.

   Una quarta línia seguiria tot el camí que obre l'estudi de la complexitat dels problemes de prova en relació a la bondat dels algorismes, donat el seu disseny. Els resultats presentats al capítol 4 indiquen un paper determinant en el procés de comparació d'algorismes per part de les mètriques de complexitat, bastant superior al que han tingut fins ara. Un primer aspecte a treballar serien els problemes en què els algorismes s'assajen sobre problemes de prova amb més de 2 classes, per als quals les mètriques de complexitat d'ús habitual no es poden utilitzar, tal i com estan definides.

5. Implementació de les propostes en una eina orientada a l'usuari.

   Finalment, una darrera possibilitat seria la implementació de totes aquestes metodologies proposades en una eina que permetés a la comunitat de ML seguir-les en cada procés d'anàlisi d'un nou algorisme que es dugui a terme. Una eina que, per exemple, impedís l'aplicació d'un test d'inferència quan no es complissin les condicions per a fer-ho o que, com a mínim, fes conscient a l'usuari d'aquest fet.

Bibliografia

[1] F. Provost, T. Fawcett, and R. Kohavi. The case against accuracy estimation for comparing induction algorithms. *Proceedings of the Fifteenth International Conference on Machine Learning - ICML 1998.*

[2] A. Fornells, E. Golobardes, J.M. Martorell, J.M. Garrell, E. Bernadó, and N. Macià. A methodology for analyzing the case retrieval from a clustered case memory. In *7th International Conference on Case-Based Reasoning*, volume 4626 of *Lecture Notes in Artificial Intelligence*, pages 122–136. Springer-Verlag, 2007.

[3] A. Asuncion and D.J. Newman. UCI machine learning repository, 2007.

[4] J. Bacardit and J. M. Garrell. Analysis and improvements of the adaptive discretization intervals knowledge representation. In *GECCO 2004: Proceedings of the Genetic and Evolutionary Computation Conference*, pages 726–738. Springer-Verlag, LNCS 3103, 2004.

[5] A. Fornells, E. Golobardes, J. M. Martorell, and J. M. Garrell. Use of SOM for estimating the probability of membership to a pattern for improving the reliability of breast cancer diagnosis. *International Journal of Neural Systems*, 2007.

[6] M. Friedman. A comparison of alternative tests of significance for the problem of m rankings. *The Annals of Mathematical Statistics*, 11:86–92, 1940.

[7] H. J. Keselman, J. Algina, and R. K. Kowalchuk. The analysis of repeated measures designs: a review. *British Journal of Mathematical and Statistical Psychology*, 54:1–20, 2001.

[8] J. L. Myers and A. D. Well. *Research design and statistical analysis.* Harper Collins Publishers, New York, 1991.

[9] J. Demsar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

[10] E. Frank and I.H. Witten. *Data mining: Practical machine learning tools and techniques with Java implementations.* Morgan Kaufman, San Francisco, CA, USA, 2000.

[11] J. Alcala-Fdez, M.J. del Jesus, J.M. Garrell, F. Herrera, C. Hervás, and L. Sánchez. *Desarrollo de una herramienta para el análisis e implementación de algoritmos de extracción de conocimiento evolutivos*, pages 413–424. Red Espa nola de Minería de Datos y Aprendizaje (TIC2002-11124-E), 2004.

[12] J. Kent Martin and Daniel S. Hirschberg. Small sample statistics for classification error rates (i): Error rate measurements. Technical Report ICS-TR-96-22, 1996.

[13] T.G. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1924, 1998.

[14] *ICML 2006, 23th International Conference on Machine Learning*, Pittsburgh (USA), 2006.

[15] *ICCBR 2005, 6th International Conference on Case-Based Reasoning*, Chicago, Illinois (USA), 2005.

[16] E. Corchado, H. Yin, and V. Bottiand C. Fyfe, editors. *Intelligent Data Engineering and Automated Learning - IDEAL 2006*, Burgos (ESP), 2006.

[17] S. Geisser. *Predictive Inference: An Introduction.* Chapmann-Hall, New York (USA), 1993.

[18] W.S.Gosset. The probable error of a mean. *Biometrika*, 6:1–25, 1908.

[19] D. J. Sheskin. *Handbook of Parametric and Nonparametric Statistical Procedures.* CRC Press, 1997.

[20] R.R. Bouckaert. Choosing between two learning algorithms based on calibrated tests. *Proceedings of the Twentieth International Conference on Machine Learning - ICML 2003.*

[21] A. Fornells, E. Golobardes, D. Vernet, and G. Corral. Unsupervised case memory organization: Analysing computational time and soft computing capabilities. In *8th European Conference on Case-Based Reasoning*, volume 4106 of *Lecture Notes in Artificial Intelligence*, pages 241–255. Springer-Verlag, 2006.

[22] A. Fornells, E. Golobardes, J. M. Martorell, J. M. Garrell, E. Bernadó, and N. Macià. Measuring the applicability of self-organization maps into case-based reasoning. *Lecture Notes in Computer Science*, 4478:532–539, 2007.

[23] N. Macià, E. Bernadó, A. Fornells, E. Golobardes, J. M. Martorell, and J. M. Garrell. Revisión sobre métricas de complejidad en el modelado de clústers de un sistema CBR. In *IV Taller de Minería de Datos, TAMIDA*, page in press, 2007.

[24] David H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992.

[25] Wray L. Buntine. Myths and legends in learning classification rules. In *Association for the Advancement of Artificial Intelligence*, pages 736–742, 1990.

[26] J. Hadamard. Sur les problèmes aux dérivés partielles et leur signification physique. *Princeton University Bulletin*, 13:49–52, 1902.

[27] C. Schaffer. A conservation law for generalization performance. In *International Conference on Machine Learning*, pages 259–265, 1994.

[28] B. Efron. Estimating the error rate of a prediction rule: improvement on crossvalidation. *Journal of the American Statistical Association*, 78:316–330, 1983.

[29] Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. *Classification and Regression Trees*. Wadsworth International Group, Belmont, CA, 1984.

[30] A. K. Jain, R. C. Dubes, and C. Chen. Bootstrap techniques for error estimation. *IEEE transactions on pattern analysis and machine intelligence*, 9:628–633, 1987.

[31] S. Weiss and N. Indurkhya. Small sample decision tree pruning. *Proceedings of the International Conference on Machine Learning - ICML 1994*, pages 335–342, 1994.

[32] R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *International Joint Conferences on Artificial Intelligence*, pages 1137–1145, 1995.

[33] R. Kohavi. Wrappersfor performance enhancement and oblivious decision graphs, 1995.

[34] D. Wolpert. The relationship between PAC, the statistical physics framework, the bayesian framework and the VC framework. In *Workshop Formal Approaches to Supervised Learning*, pages 117–214, 1994.

[35] P. Langley. Crafting papers on machine learning. In *Proc. 17th International Conf. on Machine Learning*, pages 1207–1212. Morgan Kaufmann, San Francisco, CA, 2000.

[36] N. Lachiche, C. Ferri, and S. A. Macskassy. *Proceedings of the 3rd International Workshop on ROC Analysis in Machine Learning*. Pittsburgh, 23rd International Conference on Machine Learning, 2006.

[37] E. Alpaydin. Combined 5x2 cv F test for comparing supervised classification learning algorithms. *Neural Computation*, 11:1885–1892, 1999.

[38] C. Nadeau and Y. Bengio. Inference for the generalization error. *Machine Learning*, 52:239.

[39] Geoffrey I. Webb. Multiboosting: A technique for combining boosting and wagging. *Machine Learning*, 40(2):159–196, 2000.

[40] Y. Bengio and Y. Grandvalet. No unbiased estimator of the variance of k-fold cross-validation. *Journal of Machine Learning Research*, 5:1089–1105, 2004.

[41] Remco R. Bouckaert and Eibe Frank. Evaluating the replicability of significance tests for comparing learning algorithms. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 3–12, 2004.

[42] Remco R. Bouckaert. Estimating replicability of classifier learning experiments. In *ICML '04: Proceedings of the twenty-first international conference on Machine learning*, page 15, New York, NY, USA, 2004. ACM Press.

[43] S. Salzberg. On comparing classifiers: Pitfalls to avoid and a recommended approach. *Data Mining and Knowledge Discovery*, 1(3):317–328, 1997.

[44] D. Hull. Information retrieval using statistical classification, 1995.

[45] E. G. Vázquez, A. Yañez, P. Galindo, and J. Pizarro. Repeated measures multiple comparison procedures applied to model selection in neural networks. In *IWANN '01: Proceedings of the 6th International Work-Conference on Artificial and Natural Neural Networks*, pages 88–95, London, UK, 2001. Springer-Verlag.

[46] J. Pizarro, E. Guerrero, and P. L. Galindo. Multiple comparison procedures applied to model selection. *Neurocomputing*, 48(1-4):155–173, 2002.

[47] T. K. Ho and M. Basu. Complexity measures of supervised classification problems. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(3):289–300, 2002.

[48] M. Basu and T. K. Ho. *Data Complexity in Pattern Recognition (Advanced Information and Knowledge Processing)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[49] E. Golobardes, X. Llorà, M. Salamó, and J. Martí. Computer aided diagnosis with case-based reasoning and genetic algorithms. *Journal of Knowledge Based Systems*, 15:45–52, 2002.

[50] A. Fornells. Diploma d'Estudis Avançats, 2006.

[51] J. Bacardit and J.M. Garrell. Evolving multiple discretizations with adaptive intervals for a pittsburgh rule-based learning classifier system. In *Proceedings of the Genetic and Evolutionary Computation Conference - GECCO2003*, pages 1818–1831. Lecture Notes in Computer Science 2724, Springer-Verlag, 2003.

[52] J. Ross Quinlan. *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.

[53] David W. Aha, Dennis F. Kibler, and Mark K. Albert. Instance-based learning algorithms. *Machine Learning*, 6(1):37–66, 1991.

[54] J.M. Garrell, E. Golobardes, E. Bernadó, and X. Llorà. Automatic diagnosis with genetic algorithms and case-based reasoning. *Artificial Intelligence in Engineering*, 13(4):362–367, 1999.

[55] A. Oliver, J. Freixenet, and R. Zwiggelaar. Automatic classification of breast density. *IEEE International Conference on Image Processing*, 2:1258–1261, 2005.

[56] A. Oliver, J. Freixenet, A. Bosch, D. Raba, and R. Zwiggelaar. Automatic classification of breast tissue. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 431–438, 2005.

[57] E. Golobardes, X. Llorà, J. M. Garrell, D. Vernet, and J. Bacardit. Genetic classifier system as a heuristic weighting method for a case-based classifier system. *Butlletí de l'Associació Catalana d'Intel.ligència Artificial*, 22:132–141, 2000.

[58] J. Swets. Measuring the accuracy of diagnostic systems measuring the accuracy of diagnostic systems. *Science*, 240:1285–1293, 1998.

[59] Chris Drummond and Robert C. Holte. What ROC Curves Can't Do (and Cost Curves Can). In *ROC Analysis in Artificial Intelligence*, pages 19–26, 2004.

[60] L. Breiman. Heuristics of instability and stabilization in model selection. *Ann. Statist.*, 24:2350–2383, 1996.

[61] T. Kohonen. Analysis of a simple self-organizing process. *Biological Cybernetics*, 44:135–140, 1982.

[62] S. M. Weiss and C. A. Kulikowski. *Computer Systems That Learn.* Morgan Kauffman, San Mateo, CA, 1990.

[63] F. Provost, T. Fawcett, and R. Kohavi. The case against accuracy estimation for comparing induction algorithms. *Proceedings of the Fifteenth International Conference on Machine Learning - ICML 1998.*

[64] G. J. McLachlan. The bias of the apparent error rate in discriminant analysis. *Biometrika*, 63:239–244, 1976.

[65] D. H. Wolpert. On the connection between in-sample testing and generalization error. *Sankhya: The Indian Journal of Statistics*, 52:314–345, 1992.

[66] B. Efron and R. Tibshirani. Cross-validation and the bootstrap: Estimating the error rate of a prediction rule, 1995.

[67] R. A. Olshen, E. A. Gilpin, H. Henning, M. L. LeWinter, D. Collins, and J. Ross. Twelve-month prognosis following myocardial infarction: Classication trees, logistic regression, and stepwise linear discrimination. *Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer*, 1, 1985.

[68] C. Bai. Asymptotic properties of some samples reuse methods for prediction and classication, PhD thesis. *Univesrity of California, San Diego*.

[69] J. W. Shavlik and T. G. Dietterich. General aspects of machine learning. In J. W. Shavlik and T. G. Dietterich, editors, *Readings in Machine Learning*, pages 1–10. Kaufmann, San Mateo, CA, 1990.

[70] J. R. Quinlan. Induction of decision trees. In J. W. Shavlik and T. G. Dietterich, editors, *Readings in Machine Learning*, pages 57–69. Kaufmann, San Mateo, CA, 1990.

[71] D. H. Wolpert. Stacked generalization. Technical Report LA-UR-90-3460, Los Alamos, NM, 1990.

[72] L. Breiman and P. Spector. Submodel selection and evaluation in regression: the x-random case. *Internationl Review of Statistics*, 3:291–3195, 1992.

[73] C. Cortes, L. D. Jackel, S. A. Solla, V. Vapnik, and J. S. Denker. Learning curves: Asymptotic value and rate of convergence. *Advances in Neural Information Processing Systems*, 1994.

[74] S. A. Teukolsky W. H. Press, W. T. Vetterling, and B. P. Flannery. *Numerical recipes in FORTRAN. The art of scientific computing.* Cambridge: University Press, c1992, 2nd ed., 1992.

[75] A. Fornells-Herrera, E. Golobardes-Ribé, J.M. Martorell-Rodon, and J.M. Garrell-Guiu. Are the models alone in SOM? no, they have neighbours. Presented at *Artificial Intelligence, Special Issue on Health Sciences*.

[76] A. Orriols-Puig and E. Bernadó-Mansilla. Bounding XCS parameters for unbalanced datasets. In *Proceedings of the 8th annual Conference on Genetic and Evolutionary Computation*, pages 1561–1568, 2006.

[77] B. Sierra, I. Inza, and P. Larrañaga. On applying supervised classification techniques in medicine. In *Medical Data Analysis*. Springer, 2001.

[78] T.K. Ho and M. Basu. Complexity measures of supervised classification problems. *IEEE transactions on pattern analysis and machine intelligence*, 24:289–300, 2002.

[79] T.K. Ho and E. Bernadó-Mansilla. Data complexity and domains of competence of classifiers. In M. Basu and T.K. Ho, editors, *Data Complexity in Pattern Recognition*. Springer, 2005.

[80] A. Aamodt and E. Plaza. Case-based reasoning: Foundations issues, methodological variations, and system approaches. *AI Communications*, 7:39–59, 1994.

[81] S. Haykin. *Neural Networks: A Comprehensive Foundation*. Prentice Hall, 2nd edition, 1999.

[82] M. Heath, K. Bowyer, D. Kopans, R. Moore, and P.J. Kegelmeyer. The digital database for screening mammography. *International Workshop on Digital Mammography*, 2000.

[83] J. Suckling, J. Parker, and D.R. Dance. The mammographic image analysis society digital mammogram database. In A.G. Gale et al., editor, *Proceedings of 2nd International Workshop on Digital Mammography*, pages 211–221, 1994.

[84] T. H. Samuels. *Illustrated Breast Imaging Reporting and Data System BIRADS*. American College of Radiology Publications, 3rd edition, 1998.

[85] T. Kohonen. The self-organizing map. In *Proceedings of the IEEE*, 1990.

[86] T. Honkela. Self-organizing maps in natural language processing, 1997.

[87] Jason Eisner. State-of-the-art algorithms for minimum spanning trees: A tutorial discussion. 1997.

[88] F. Lebourgeois and H. Emptoz. Pretopological approach for supervised learning. *International Conference on Pattern Recognition*, 04:256, 1996.

[89] I.Y. Kim and O.L. de Weck. Adaptive weighted-sum method for bi-objective optimization: Pareto front generation. *Struct Multidisc Optim*, 29:149–158, 2005.

[90] John A. Rafter, Martha L. Abell, and James P. Braselton. Multiple comparison methods for means. *SIAM Review*, 44(2):259–278.

[91] Jason C. Hsu. *Multiple Comparisons. Theory and methods*. Chapman and Hall, 1996.

[92] M. Hollander and D. A. Wolfe. *Nonparametric statistical methods (2nd Edition)*. Wiley, New York, 1991.

[93] Harald Cramer. *Mathematical Metods of Statistics*. Princeton University Press, 1964.

[94] D. Howell. *Statistical Methods for Psychology*. Wadsworth, 2006.

[95] A. Leon-García. *Probability and Random Processes for Electrical Engineering*. Addison-Wesley, 1994.

[96] Donald E. Knuth. *The Art of Computer Programming, Volume II: Seminumerical Algorithms, 2nd Edition*. Addison-Wesley, 1981.

[97] H. Abdi and P. Molin. Lilliefors test of normality. In *Encyclopedia of Measurement and Statistics*. Thousand Oaks, 2007.

[98] G. G. Judge, R. C. Hill, W. E. Griffiths, H. Lütkepohl, and T. Lee. *Introduction to the Theory and Practice of Econometrics*. John Wiley, New York, 1982.

[99] D. N. Gujarati. *Basic Econometrics*. Mcgraw-Hill, 2002.

[100] C. M. Jarque and A. K. Bera. Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters*, 6(3):255–259, 1980.

[101] S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52:591–611, 1965.

[102] T. W. Anderson and D. A. Darling. Asymptotic theory of certain "goodness-of-fit" criteria based on stochastic processes. *Annals of Mathematical Statistics*, 23:193–212, 1952.

[103] M. A. Stephens. Edf statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association*, 69:730–737, 1974.

[104] G. W. Snedecor and W. G. Cochran. *Statistical Methods*. Iowa State University Press, 1989.

[105] M. B. Brown and A. B. Forsythe. Robust tests for the equality of variances. *Journal of the American Statistical Association*, 69:364–367, 1974.

[106] R. L. Kirk. *Experimental designs: Procedures for the behavioral sciences*. Brooks/Cole Publishing Company, 1982.

[107] J. Cohen. *Statistical power analysis for the behavioral sciences.* Academic Press, 1988.

[108] L. A. Marascuilo and M. McSweeney. *Nonparametric and distribution-free methods for the social sciences.* Brooks/Cole Publishing Company, 1977.

[109] W. W. Daniel. *Applied nonparametric statistics.* PWS-Kent Publishing Company, 1990.

[110] L. A. Marascuilo and R. C. Serlin. *Statistical Methods for the social and behavioral sciences.* W.H.Freeman and Company, 1988.

[111] E. Martínez et. al. Morphological analysis of mammary biopsy images. In *Proceedings of the IEEE International Conference on Image Processing*, 1996.

[112] J. Martí et. al. Shape-based feature selection for microcalcification evaluation. In *Imaging Conference on Image Processing*, 1998.

[113] S. E. Maxwell and H. D. Delaney. *Designing experiments and analizing data.* Waldsworth Publishing Company, Belmont, CA, 1990.

[114] D. C. Howell. *Statistical Methods for psychology.* PWS-Kent Publishing Company, Boston, 1992.

[115] R. A. Fisher. *Statistical methods and scientific inference.* Hafner Publishing Co., New York, 1959.

[116] L. C. Hamilton. *Modern data Analysis: A First Course in Applied Statistics.* Wadsworth, Belmont, CA, 1990.

[117] G. Keppel. *Design and Analysis: A researcher's handbook.* Prentice Hall, Englewood Cliffs, NJ, 1991.

[118] S. W. Greenhouse and S. Geisser. On methods in the analysis of profile data. *Psychometrika*, 24:95–112, 1959.

[119] H. Huynh and L. S. Feldt. Conditions under which mean square ratios in repeated measurements designs have exact f-distributions. *Journal of the American Statistical Association*, 65(32):1582–1589, 1970.

[120] O. J. Dunn. Multiple comparisons among means. *Journal of the American Statistical Association*, 56:52–64, 1961.

[121] J. W. Tukey. *The problem of multiple comparisons*. Princeton University Press, Princeton, NJ, 1953.

[122] J. B. Winer, D. R. Brown, and K. M. Michels. *Statistical principles in experimental design*. McGraw-Hill Publishing Company, New York, 1991.

[123] M. Keuls. The use of studentized range in connection with an analysis of variance. *Euphytica*, 1:112–122, 1952.

[124] M. Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of American Statistical Association*, 32:675–701, 1937.

[125] M. Friedman. A correction: The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of American Statistical Association*, 34:109, 1939.

[126] R. L. Iman and J. M. Davenport. Approximations of the critical region of the friedman statistic. *Communications in Statistics*, pages 571–595, 1980.

[127] J. H. Zar. *Biostatistical Analysis (4th ed.)*. Prentice Hall, New Jersey, 1998.

[128] W. J. Conover. *Practical nonparametric statistics (2nd ed.)*. John Wiley and Sons, New York, 1980.

[129] E. B. Page. Ordered hypothesis for multiple treatments: a significance test for linear ranks. *Journal of American Statistical Association*, 58:216–230, 1963.

[130] P. B. Nemenyi. Distribution-free multiple comparisons.

[131] S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6:65–70, 1979.

[132] G. Hommel. A stagewise rejective multiple test procedure based on a modified test. *Biometrika*, 75:383–386, 1988.

[133] Y. Hochberg. A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, 75:800–803, 1988.

[134] B. Holland. On the application of three modified boferroni procedures to pairwse multiple comparisons in balanced repeated measures designs. *Computational Statistics Quarterly*, 6:219–231, 1991.

[135] J. von Neumann. Various techniques used in connection with random digits. monte carlo methods. *Nat. Bureau Standards*, 12:36–38, 1951.

[136] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods (2nd ed.)*. Springer-Verlag, New York, 2004.

[137] J. H. Holland. *Adaptation in Natural anb Artificial Systems*. University of Michigan Press, 1975.

[138] S. W. Wilson. Classifier fitness based on accuracy. *Evolutionary Computation*, 3:149–175, 1995.

[139] X. Llorà and J.M. Garrell. Knowledge-independent data mining with fine-grained parallel evolutionary algorithms. In *Proceedings of the Third Genetic and Evolutionary Computation Conference*, pages 461–468. Morgan Kaufmann, 2001.

[140] K. A. DeJong and W. M. Spears. Learning concept classification rules using genetic. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 651–656, 1991.

[141] D. Winfields, M. Silbiger, and G. Brown. Technology transfer in digital mamography. In *Report of the Joint National Cancer Institute, Workshop of May 19-20, Invest Radiology*, pages 507–515, 1994.

[142] A. Fornells, E. Golobardes, X. Vilasis, and J. Martí. Integration of strategies based on relevance feedback into a tool for retrieval of mammographic images. *Lecture Notes in Computer Science*, pages 116–124, 2006.